Wavelet-Based Methodology in Data Mining for Functional Data

A Thesis

Presented to

The Academic Faculty

by

Myong-Kee Jeong

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy in Industrial and Systems Engineering

Georgia Institute of Technology

May 2004

Wavelet-Based Methodology in Data Mining for Complicated Functional Data

Approved:

Jye-Chyi Lu, Chairman

Kwok-Leung Tsui

Brani Vidakovic

Xiaoming Huo

David W. Rosen

Date Approved: March 26, 2004

Acknowledgements

First, I give thanks to God for giving me new life and vision in Him. He has given me a strength and encouragement. I have to confess that my life is meaningless without Him.

I would like to express my sincere gratitude to my advisor, Dr. Jye-Chyi Lu, for his continuous support, warm encouragement, and invaluable guidance throughout my doctoral program. His unique insight, enthusiasm and prudent guidance and advice have been truly inspirational. He has been a most enthusiastic research supervisor.

I would like to extend my appreciation to Dr. Brani Vidakovic, who gave me insights into my research on the applications of wavelets. I am greatly enlightened by his valuable suggestions for the improvement of my dissertation.

I would also like to thank Dr. Kwok-Leung Tsui, Dr. David Rosen, and Dr. Xiaoming Huo for serving on my doctoral committee and for their valuable input and guidance during my doctoral studies.

I am very grateful to Dr. Chen Zhou, Dr. Paul Kvam, Dr. Roshan Vengazhiyil, and Dr. Seong-Hee Kim for their academic support. They have been of great help to me throughout my years at Georgia Tech.

I would like also to extend my gratitude to the entire Department of Industrial and Systems Engineering at the Georgia Institute of Technology for its high standards in both education and support for students.

I thank lab members, Uk and Hyoungtae, for their friendship. I would like to thank Pastor Sunghee Lim and Pastor Bongsoo Choi for their prayers. I also want to thank all the members of the Friday Bible Study at GNBC church

Finally, I wholeheartedly thank my wife, Jeonghee, and my precious children, Heymin and Grace, for their prayers and support. I also thank my parents, parents-in-law, sisters and brother, all of whom have been encouraged me throughout my years at Georgia Tech.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Summary

To handle potentially large size and complicated nonstationary functional data, we present the wavelet-based methodology in data mining for process monitoring and fault classification.

Since traditional wavelet shrinkage methods for data de-noising are ineffective for the more demanding data reduction goals, this thesis presents data reduction methods based on discrete wavelet transform. Our new methods minimize objective functions to balance the tradeoff between data reduction and modeling accuracy. An upper bound of a data signal's approximation (or estimation) error is derived. Several evaluation studies with four popular testing curves used in the literature and with two real-life data sets demonstrate the superiority of the proposed methods to engineering data compression and statistical data de-noising methods that are currently used to achieve data reduction goals.

Further experimentation in applying a classification tree-based data mining procedure to the reduced-size data to identify process fault classes also demonstrates the excellence of the proposed methods. In this application the proposed methods, compared with analysis of original large-size data, result in lower misclassification rates with much better computational efficiency.

One deficiency in the procedures developed from the wavelet coefficients provided from the discrete wavelet transform (DWT) is the lack of shift-invariance. If the signal to be analyzed is shifted even by a small amount, the corresponding wavelet transform coefficients do not experience the same simple translation. Instead, the coefficients are modified in a

much more complex manner, due to the fact that the WT is critically sampled. Direct assessment of the wavelet coefficients can lead to inaccurate decisions for fault detection against time-shift. A scale-wise energy representation such as a scalogram provides a more robust signal feature for fault detection against time-shift than the DWT coefficients directly. This thesis extends the scalogram's ability for handling noisy and possibly massive data which show time-shifted patterns. The proposed thresholded scalogram is built on the fast wavelet transform, which can effectively and efficiently capture non-stationary changes in data patterns. The asymptotic distribution of the thresholded scalogram is derived. This leads to large sample confidence intervals that are useful in detecting process faults statistically, based on scalogram signatures. Application of the scalogram-based data mining procedure (mainly, Classification and Regression Trees) demonstrates the potential of the proposed methods for use in analyzing complicated signals to arrive at engineering decisions.

Using the special ability of the DWT in modeling sharp-change data, we present several statistical process control charting (SPC) procedures for functional data. Some available approaches use threholding methods or engineering knowledge to select and specify the wavelet coefficients that will be monitored. Because these approaches fix the wavelet coefficients to be monitored, they are not sensitive new types of faults. In this thesis, we present a SPC procedure that adaptively determines which wavelet coefficients will be monitored, based on their shift information, which is estimated from process data. By adaptively monitoring the process, we can improve the performance of the control charts for functional data. Using a simulation study, we compare the performance of some of the recommended approaches.

CHAPTER I

INTRODUCTION

1.1 Motivation

Advanced technology such as various types of automatic data acquisition, information management, and systems for communication networking has created a tremendous capability for managers to access valuable information from various manufacturing enterprise to improve their operational quality and efficiency. Data mining and signal processing techniques are more popular than ever in many fields including intelligent manufacturing. As data sets increase in size, their exploration, manipulation, and analysis become more complicated and consume more resource. Timely synthesized information is needed for product design, process trouble-shooting, quality/efficiency improvement and resource allocation decisions.

To our knowledge, most of the successful data mining applications with large size data have been in the sale of groceries and fashion goods, management of customer relations, and analysis of telecommunications fraud and a few other fields outside of the manufacturing or process control arenas. In this thesis, we present a wavelet-based methodology for mining functional data for use in process monitoring and fault classification. Many types of functional data are available in industrial processes for process monitoring, control and fault-classification. Ganesan et al. (2002) presented the example of acoustic emission signals from a nano-machining process. Rying (2001) gave an example based on the control signals in an ultra-thin film chemical deposition process. Lawless et al. (1999) presented

examples from automotive engineering. Jin and Shi (2001) described the several kinds of functional data present in a stamping process. Many researchers who have attempted to use functional data in controlling or monitoring manufacturing practices, e.g., Bakshi (1999) and Ganesan *et al.* (2002), have encountered difficulties in handling complicated functional data with nonstationary, correlated or dynamically changing patterns that are contributed by potential process faults.

Many researchers have recommended wavelet-based methods to handle this type of nonstationary and possibly correlated data. Discrete wavelet transforms (DWT) are better able to model irregular data patterns than the Fourier transform and standard statistical procedures, e.g., splines, polynomial and nonparametric regressions, and provide a multiresolution approximation to the data (Mallat, 1989). The usefulness of wavelet transforms have been demonstrated in image and audio compression practices (e.g., Rao and Bopardikar, 1998; Chapter 5) and in many data-denoising studies (e.g., Donoho and Johnstone, 1994) across various applications. Thus, we will focus on wavelet-based data reduction procedures for complicated functional data.

The aim of our data reduction is to produce a small set of representative data suitable for many kinds of decisions. Moreover, if it is necessary, an accurate approximation of the original data could be obtained for many types of analysis, i.e., our procedure has the properties of data compression. Thus, the underlying theme of our methodology is to reduce within a mathematically rigorous framework the size of data and apply existing and new procedures to this reduced-size data for various decision-making purposes.

In many processes, the quality of a process is characterized by a functional data, which is a time-sequence data. The process monitoring of functional data is difficult because

it is highly dimensional and nonstationary. A general framework for monitoring wavelet coefficients has yet to be developed. A few existing approaches construct an appropriate test statistic based on the reduced-size data. However, the dimension of the reduced-size data can be still high. Moreover, because the wavelet coefficients to be monitored are specified, they are insensitive to new types of fault. In this thesis, we present a statistical process control (SPC) procedure that adaptively determines which wavelet coefficients to monitor based on their shift information, which is estimated from the process data. By adaptively monitoring the process, we can improve the performance of the control charts for functional data.

The lack of shift-invariance is one deficiency in the procedures developed on the basis of the wavelet coefficients provided from the DWT. If the signal to be analyzed is shifted even slightly, the corresponding wavelet transform coefficients do not experience the same simple translation. Instead, the coefficients are modified in a much more complex manner because the WT is critically sampled. Scalograms provide measures of signal energy at various frequency bands and are commonly used to make decisions in many fields including signal and image processing, astronomy and metrology. Scalograms provide a more robust signal feature for fault detection involving time shifts than the DWT coefficients directly. In estimating a signal's functional pattern with noisy data, Donoho and Johnstone (1994) proposed a data denoising procedure based on the idea of thresholding out secondary wavelet coefficients representing data noises. In many applications in data mining, the large size of non-stationary data makes computations inefficient (see Pittner and Kamarthi 1999 for an example). Extending the usefulness of the popular scalogram to noisy and possibly massive data, this thesis develops a thresholded scalogram and studies its properties and

applicability to engineering decision making.

1.2 Thesis Outline

This thesis is organized as follows. Chapter 2 presents data reduction methods based on discrete wavelet transform to handle potentially large sized and complicated nonstationary functional data. An upper bound of data signal's approximation (or estimation) error is derived. Based on evaluation studies with popular testing curves and real-life data sets, the proposed methods demonstrate their competitiveness to the existing engineering data-compression and statistical data-denoising methods for achieving the data reduction goals. Further experimentation of applying a classification tree-based data mining procedure to the reduced-size data for identifying process fault classes illustrate the potential of the proposed ideas compared with analysis of original larger-size data.

Chapter 3 provides statistical process control charting (SPC) procedures to monitor the process with nonstationary and time-dependent functional data. Our proposed SPC procedure adaptively determines which wavelet coefficients to monitor. This determination is based on their shift information, which is estimated from the process data. Using a simulation study, we compare the performance of our proposed procedures with some of the recommended approaches. A real-life example is provided.

Chapter 4 develops a thresholded scalogram and studies its properties and applicability to engineering decision making by extending the usefulness of the popular scalogram to noisy and possibly massive data. The proposed thresholded scalogram is built on the fast wavelet transform, which can capture non-stationary changes in data patterns effectively and efficiently. The asymptotic distribution of the thresholded scalogram is derived. This leads

to large sample confidence intervals that are useful in detecting process faults statistically, based on scalogram signatures. Application of the scalogram-based data mining procedure (mainly, Classification and Regression Trees) is provided to demonstrate the potential of the proposed methods for analyzing complicated signals as a basis for engineering decisions.

Finally, Chapter 5 summarizes research results and proposes research problems for future investigation.

CHAPTER II

WAVELET-BASED DATA REDUCTION TECHNIQUES

Traditionally used process fault detection procedures have difficulty locating local changes effectively for processes with a large size of nonstationary data. This chapter proposes wavelet-based methods for reducing complicated large size data to facilitate the possibility of using well-known decision-making procedures for identifying process changes and classifying process fault types.

2.1 Introduction

There are many types of large size data. The data studied in this thesis do not have many attributes (e.g., data from grocery sales) for "dimension reduction." We focus on the the data with complicated nonstationary patterns. Figure 1 presents an example of data taken from Nortel's wireless antenna manufacturing processes. There are more than 30,000 data points in one atnenna data set with complicated patterns. Timely synthesized information was needed for product design validation, process trouble shooting and production quality improvement. However, the local changes in the cusps and lobes of the data were difficult to handle for traditional data analysis tools. This motivates the focus of this article: *developing general-purpose data-reduction procedures for commonly used data analysis tools to be useful in handling large-size complicated functional data.* See Ganesan, Das, Sikdar and Kumar

(2003) for another motivating example from nano-manufacturing processes.

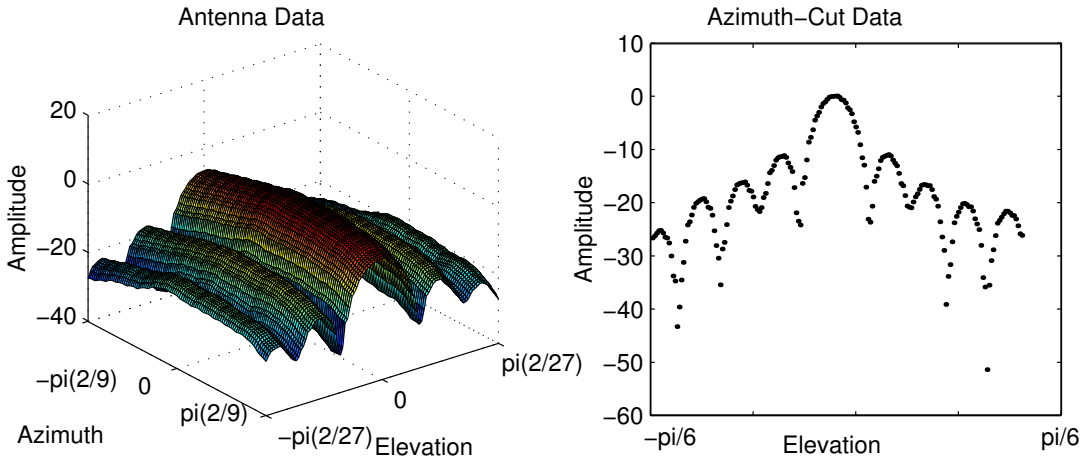

Figure 1: Data Signals from Antenna Manufacturing Processes

Several data-reduction procedures are available in the literature. Lu (2001) summaized them into three main categories: sampling approaches, modeling and transformation techniques, and data splitting methods. Even with these methods, it is recognized that complicated functional or spatial data with nonstationary, corrleated or dynamically changing patterns contributed from potential process faults are difficult to handle. Wavelet transforms model irregular data patterns such as cups and lobes in Figure 1 better than the Fourier transform and standard statistical procedures, e.g., splines and polynomial regressions, and provide a multi-resolution approximation to the data (Mallat, 1998). Applications of wavelet-based procedures in solving manufacturing problems include: Jin and Shi (1999) used tonnage signals to detect faults in a sheet-metal stamping process; Wang, Chen, Yang, and McGreavy (1999) used different catalyst recycling rates to diagnose failures in a residual fluid catalytic cracking process; and Lada, Lu and Wilson (2002) analyzed quadrupole mass spectrometry (QMS) samples of a rapid thermal chemical vapor deposition (RTCVD) process to detect significant deviations from the nominal processes.

Using the knowledge of experts from a particular process, one could derive a "feature-preserving" procedure (Jin and Shi, 1999) to extract a particular data pattern represented by a set of a few "reduced-size" data. Then, link them to a specific type of process fault for monitoring production performance. More rigorously, if the "reduced-size data set" is constructed to detect specific types of known faults, a data-reduction procedure could be derived to minimize Type-I and/or -II errors in hypothesis testing of the occurrence of faults. For example, Jin and Shi's (2001) optimal number of wavelet coefficients included in the fault classification is based on the minimization of probabilities of misclassification errors using SPC limits as the decision rule. However, the wavelet coefficients selected for a given decision rule might not be suitable for other purposes of analysis, e.g., fault classification, failure prediction, and data clustering to improve manufacturing quality and efficiency. Thus, the aim of our data-reduction is to produce a small set of "representative data" suitable for various data and decision analyses either planned or unplanned before seeing the data.

Data-denoising procedures such as *VisuShrink* (Donoho and Johnstone, 1994) and *RiskShrink* (Donoho and Johnstone, 1995) are used as data-reduction tools in several applications, e.g., Jin and Shi (2001), Ganesan *et al.* (2003). Rying, Gyurcsik, Lu, Bilbro, Parsons, and Sorrell (1997) applied a scale-dependent energy metric, E_s = sum of squares of all wavelet coefficients at atoms $\phi_{s,u}$ across all u positions at the same scale s, to the Ar^+ signals in a semiconductor fabrication experiment. The scalogram (Vidakovic 1999, page 289) plots these energy metrics at different resolution scales for visualizing the data-energy distribution. These energy metrics can served as representative reduced-size data such that procedures such as the linear discriminant analysis method can detect and distinguish process faults

timely.

The purposes of data-denoising and data-reduction are different. Data in engineering applications do not have large-size random noises for showing the effectiveness of data-denoising procedures. On the other hand, the energy-metric approach is too aggressive and not linked to local data characteristics. For example, any functional curve with 1,024 data points will have only six E_s measures.

This chapter develops a well motivated objective function for selecting the reduced-size data, derives the "thresholding parameter" to optimize the objective function, and evaluates the properties of the data-reduction procedures with several simulation experiments and real-life data analyses.

2.2 Wavelet Transforms

Wavelets are fundamental building block functions, analogous to the sine and cosine functions of the Fourier transform. Wavelets are localized basis functions that are translated and dilated versions of some fixed mother wavelet. Some signals often show a non-stationary and transient nature and carry small yet informative components embedded in larger repetitive signals. Wavelets have flexible time-frequency resolution and make up an efficient alternative for use in quantifying such transient signals.

A wavelet is a function $\psi(t) \in L^2(\mathbb{R})$ with the following basic properties

$$\int_{\mathbb{R}} \psi(t)\, dt = 0 \quad \text{and} \quad \int_{\mathbb{R}} \psi^2(t)\, dt = 1,$$

where $L^2(\mathbb{R})$ is the space of square integrable real functions defined on the real line $\mathbb{R}$. Wavelets can be used to create a family of time-frequency atoms, $\psi_{s,u}(t) = s^{1/2}\psi(st-u)$, via

the dilation factor s and the translation u. We also require a scaling function $\phi(t) \in L^2(\mathbb{R})$ that satisfies

$$\int_{\mathbb{R}} \phi(t)\, dt \neq 0 \quad \text{and} \quad \int_{\mathbb{R}} \phi^2(t)\, dt = 1.$$

Selecting the scaling and wavelet functions as $\{\phi_{L,k}(t) = 2^{L/2}\phi(2^L t - k); k \in \mathbb{Z}\}$, $\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k); j \geq L, k \in \mathbb{Z}\}$, respectively, one can form an orthonormal basis to represent a signal function $f(t) \in L^2(\mathbb{R})$ as follows.

$$f(t) = \sum_{k\in\mathbb{Z}} c_{L,k}\phi_{L,k}(t) + \sum_{j\geq L}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)$$

where $\mathbb{Z}$ denote the set of all integers $\{0, \pm 1, \pm 2, \ldots\}$, and the coefficients $c_{L,k} = \int_{\mathbb{R}} f(t)\phi_{L,k}(t)\, dt$ are considered to be the coarser-level coefficients characterizing smoother data patterns, and $d_{j,k} = \int_{\mathbb{R}} f(t)\psi_{j,k}(t)\, dt$ are viewed as the finer-level coefficients describing (local) details of data patterns. In practice, the following finite version of the wavelet series approximation is used:

$$\tilde{f}(t) = \sum_{k\in\mathbb{Z}} c_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{J}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)\,, \tag{1}$$

here $J > L$ and L correspond to the coarsest resolution level.

Consider a sequence of data $\mathbf{y} = (y(t_1), \cdots, y(t_N))'$ taken from $f(t)$ or obtained as a realization of $y(t) = f(t) + \epsilon_t$ at equally spaced discrete time points $t = t_i$'s, where ϵ_{t_i}'s are independent and identically distributed (i.i.d.) noises. The discrete wavelet transform (DWT) of $\boldsymbol{y}$ is defined as

$$\boldsymbol{d} = \boldsymbol{W}\boldsymbol{y}, \tag{2}$$

where

$$\boldsymbol{W} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1N} \\ h_{21} & h_{22} & \cdots & h_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{NN} \end{pmatrix} \tag{3}$$

is the orthonormal $N \times N$ wavelet transform matrix. The matrix $\boldsymbol{W}$ is different according to the wavelet type, the decomposition level, and the number of sample points N. The elements (h_{jk}) have a special structure, corresponding to a sequence of linear filtering operations. In practice, the pyramid algorithm is used to compute the wavelet and inverse wavelet transforms in $O(N)$ operations (Mallat, 1989).

In the equation (25), let

$$\boldsymbol{d} = (\boldsymbol{c}_l, \boldsymbol{d}_l, \boldsymbol{d}_{l+1}, \cdots, \boldsymbol{d}_J)^\top, \tag{4}$$

where $\boldsymbol{c}_l = (c_{l,0}, \cdots, c_{l,2^l-1})^\top, \boldsymbol{d}_l = (d_{l,0}, \cdots, d_{l,2^L-1})^\top, \cdots, \boldsymbol{d}_J = (d_{J,0}, \cdots, d_{J,2^J-1})^\top$ are wavelet coefficients at various scales or subbands. The total number of wavelet coefficients equals the number of signal measurements, i.e., $N = 2^{J+1}$. The $c_{L,k}$'s capture the low frequency oscillations, while $d_{j,k}$'s capture the high frequency oscillations. The coefficients $d_{J,k}$'s represent the finest scale (details) and the $c_{L,k}$'s represent the coarsest scale (smooth) (Morettin 1997). To simplify the notation, we use $\boldsymbol{d} = (d_1, d_2, \ldots, d_N)^\top$ instead of using c_{Lk}, d_{jk} for the components of $\boldsymbol{d}$ without any confusing.

The computational efficiency of DWT is better than the other transforms. For example, the principal component analysis (PCA) requires solving an eigenvalue system which is an expensive $O(N^3)$ operation. The FFT requires $O(N \log N)$ operations, but a fast wavelet

transform DWT only requires $O(N)$ operations.

Using the inverse DWT, the $N \times 1$ vector $\mathbf{y}$ of the original signal curve can be "reconstructed" as $\mathbf{y} = \boldsymbol{W}'\mathbf{d}$. The process of transforming a data set via the DWT closely resembles the process of computing the Fast Fourier Transformation (FFT) of that data set.

Wavelets exist in an abundant variety, and the fundamental problem to overcome lies in deciding which wavelet will produce the best results for a particular application such as classification, clustering, process monitoring, etc. Depending on the application, we may need to choose a wavelet that satisfies special properties. Smoothness is closely related to how many times a wavelet can be differentiated and to the number of vanishing moments. A wavelet has M vanishing moments if

$$\int t^q \psi(t) dt = 0, \quad q = 0, 1, \cdots, M-1.$$

A wavelet is M times differentiable only if $\psi(t)$ has M vanishing moments.

A high amplitude wavelet coefficient occurs when the wavelet has a support that overlaps a sharp transition. The number of high wavelet coefficients created by the singularity depends on the support size of ψ, which should thus be as small as possible. Over smooth regions, the wavelet coefficients are small at fine scales if ψ has enough vanishing moments to take advantage of the large Lipschitz regularity α. However, the support size of ψ increases proportionally to the number of vanishing moments. The choice of an optimal wavelet is therefore a trade-off between the number of vanishing moments and its support size (Mallat 1998, page 519). If f has few isolated singularities and is very regular (smooth) between singularities, we must choose a wavelet with many vanishing moments to produce a large number of small wavelet coefficients. If the density of singularities increases, it might be

better to decrease the size of its support at the cost of reducing the number of vanishing moments. Indeed wavelets that overlap the singularities create high amplitude coefficients.

Thus, the general guideline for the selection of wavelet type is as follows: For smooth signals, we should use the wavelet that has a higher number of vanishing moments. The larger the vanishing moments, the fewer significant wavelet coefficients are necessary for representation. For example, the Haar wavelet is not well adapted to approximating smooth functions because it has only one vanishing moment (=a lack of smoothness). For signals with many singularities, e.g., Bumps, Blocks, we should use the wavelet that has a smaller support area, i.e., a lower number of vanishing moments.

2.3 Data Compression, Reduction and De-noising Methods

2.3.1 Signal Approximation and Data Compression Methods

In the signal processing field, linear approximation method (see Mallat (1998, Section 9.1) for details) uses the following function with a set of pre-determined vectors $\boldsymbol{g}_m$, $m = 0, 1, \ldots, M-1$, to reconstruct the original data signals,

$$\boldsymbol{f}_M = \sum_{m=0}^{M-1} < \boldsymbol{f}, \boldsymbol{g}_m > \boldsymbol{g}_m, \tag{5}$$

where $< \boldsymbol{f}, \boldsymbol{g}_m >$ is the inner product of the function $\boldsymbol{f}$ and the projected vector $\boldsymbol{g}_m$. In the wavelet-based approximation, $< \boldsymbol{f}, \boldsymbol{g}_m >$ is the wavelet coefficient (from the coarsest level to the finest level in the linear method). The nonlinear approximation method (Mallat (1998, Section 9.2)) selects the M projection vectors adaptively (e.g., M-largest wavelet coefficients (in absolute values)) using the data signal information to improve the approximation error. In both linear and nonlinear approximation methods, M is fixed by the decision-maker, or

by the pre-determined error bound, e.g., $\epsilon(M) = \sum_{i=1}^{N}[f(t_i) - f_M(t_i)]^2/N$.

The wavelet coefficients selected from the above approximation methods are usually treated as "compressed data" for reconstructing the original data signals. In this article, they are treated as "reduced-size" data in process fault detection, classification and other decisions for improving process quality.

There were very limited studies in the literature for deciding the number (M) of vectors used in the model $\boldsymbol{f}_M$ adaptively based on signal characteristics. The following presents the AMDL (Approximate Minimum Description Length) method proposed by Saito (1994). The AMDL selects M to minimize the following cost function:

$$\text{AMDL}(M) = 1.5M \log_2 N + 0.5N \log_2 \left[\sum_{i=1}^{N} (y_i - \widehat{y}_{i,M})^2 \right],$$

where $\widehat{y}_{i,M}$ is the approximation model constructed from the M largest-magnitude wavelet coefficients, and the data y_i is equal to $y(t_i) = f(t_i) + \epsilon_{t_i}$ with random normal(0, σ^2) errors. As addressed in Antoniadis *et al.* (1997), the AMDL function is similar to the Akaike information quantity commonly used in many statistical model selection procedures, including linear regression models. There are several similar model selection methods in the signal processing literature based on cost functions related to quantities defined in "information theory," e.g., entropy or mutual information (see Ihara (1993); Liu and Ling (1999) for examples).

2.3.2 Data De-noising: Shrinkage Methods

Donoho and Johnstone (1995) developed several wavelet based "shrinkage" techniques in the nonparametric regression format to find a smooth estimate ($\hat{\boldsymbol{f}}$) of $\boldsymbol{f}$ from the "noisy" data, $\boldsymbol{y}$. The performance of the estimator is measured by the expected mean square error,

$E[\sum_{i=1}^{N}(y_i - \hat{f}(t_i))^2/N]$.

By applying the DWT to the data y_i's, $\mathbf{d} = \boldsymbol{W}\boldsymbol{y}$, we obtain the following model in the wavelet domain: $d_{j,k} = \theta_{j,k}+\eta_{j,k}$, for $j = L, \cdots, J$, $k = 0, 1, \cdots, 2^j-1$, and $c_{L,k} = \theta_{L,k}+\eta_{L,k}$, for $k = 0, 1, \cdots, 2^L - 1$, where $J = \log_2 N - 1$. The model can be represented in the vector format as follows.

$$\mathbf{d} = \boldsymbol{\theta} + \boldsymbol{\eta} \tag{6}$$

where $\mathbf{d}, \boldsymbol{\theta}$ and $\boldsymbol{\eta}$ represent the collection of all coefficients, parameters and errors, respectively. Since $\boldsymbol{W}$ is an orthonormal transform, $\eta_{j,k}$'s are still i.i.d. $N(0, \sigma^2)$ (Vidakovic 1999, page 169). To simplify the notation, we use $\boldsymbol{d} = (d_1, d_2, \ldots, d_N)^\top$ instead of using c_{Lk}, d_{jk} for the components of $\boldsymbol{d}$ without any confusing.

Donoho and Johnstone (1995) developed several wavelet-based "shrinkage" techniques in the nonparametric regression to find a smooth estimate ($\hat{\boldsymbol{f}}$) of $\boldsymbol{f}$ from the "noisy" data, $\boldsymbol{y}$. In particular, their hard-thresholding policy finds the estimate of θ_i to minimize the objective function

$$\sum_{i=1}^{N}(d_i - \theta_i)^2 + \tau^2 \sum_{i=1}^{N} |\theta_i|_0 \tag{7}$$

where $\sum_{i=1}^{N} |\theta_i|_0$ is the number of non-zero coefficients selected to estimate the underlying function $\boldsymbol{f}$ (using $\hat{\boldsymbol{f}} = \boldsymbol{W}^{-1}\hat{\boldsymbol{\theta}}$). The optimal estimate $\hat{\theta}_i$ is found to be equal to d_i if $|d_i| > \tau$; otherwise, $\hat{\theta}_i = 0$. Although the parameter τ was not set as the threshold originally, it becomes the threshold in the estimate of θ_i through the minimization process.

In the shrinkage scheme, wavelet coefficients are set to zero if their absolute values are

below a certain threshold level, $\lambda > 0$. Under this scheme, we have:

$$\hat{\theta}_i = \left\{ \begin{array}{ll} \hat{d}_{h,i}(\lambda) & = d_i I(|d_i| > \lambda) \text{ (hard thresholding)}, \\ \hat{d}_{s,i}(\lambda) & = \text{sign}(d_i)\max(0, |d_i| - \lambda) \text{ (soft thresholding)}. \end{array} \right\}$$

where sign(d) satisfies sign$(d) = -1$, if $d < 0$; sign$(d) = 0$, if $d = 0$; otherwise, sign$(d) = 1$. Hard thresholding is a "keep" or "kill" rule, while the soft thresholding is a "shrink" or "kill" rule. It has been shown that hard thresholding results in larger variance while soft thresholding has larger bias. Hard thesholding is also very sensitive to small changes in the data. Soft thresholding has various advantages such as continuity of the shrinkage rule. See Bruce and Gao (1996) for a comparison study between these two thresholding policies.

By thresholding the wavelet coefficients d_i to produce $\hat{\theta}_i$, one can obtain an estimate of $\boldsymbol{f}$ from $\hat{\boldsymbol{f}} = \boldsymbol{W}^{-1}\hat{\boldsymbol{\theta}}$. Because smaller size of coefficients usually are contributed from data noises, thresholding out these coefficients has an effect of "removing data noises." Thus, the shrinkage methods are called data de-noising methods.

In using any type of wavelet thresholding, the main issue is how to choose the threshold value. Choosing a very large threshold will make it difficult for a coefficient to be included in the data signal reconstruction, consequently resulting in an oversmoothing of the data curve. On the other hand, choosing a very small threshold value will allow many coefficients to be included in the reconstruction, giving a result close to the original noisy signal. The proper choice of threshold involves a careful balance of these principles. Comprehensive overview for threshold selection is given in Antoniadis, Gijbels and Grégoire (1997). We will briefly review the following three well-known methods without giving their technical details. See references therein for their derivations.

VisuShrink uses the soft-thresholding version of the universal thresholding method

(Donoho and Johnstone, 1994). The universal thresholding value of $VisuShrink$ is $(2\ln N)^{1/2}\sigma$. This is based on the result that when ϵ_i's are a white noise sequence of independent and identically distributed $N(0,1)$ errors, then $\Pr\{\max_{1\leq i\leq N}|\epsilon_i| > \sqrt{2\log N}\} \to 0$ as $N \to \infty$. The feature of $VisuShrink$ is that it guarantees a "noise free" reconstruction, but in doing so it often underfits the data by setting the threshold conservatively high. It pays attention to smoothness rather than to minimizing the mean square error. The reconstruction will include fewer coefficients, resulting in an estimate that is much smoother than the estimate obtained from the following minimax method.

To reduce the modeling error of the above method, Donoho and Johnstone (1995) developed the $RiskShrink$. This method uses the soft-thresholding version of the minimax estimate. The optimal minimax threshold, λ_N^* can be obtained as

$\lambda_N^* =$ the largest λ attaining Λ_N^* below.

A constant Λ_N^* is calculated as a minimax quantity:

$$\Lambda_N^* \equiv \inf_{\lambda} \sup_{\mu} \frac{\rho(\lambda,\mu)}{\frac{1}{N} + \min(\mu^2, 1)}. \tag{8}$$

The function $\rho(\lambda,\mu)$ is defined as $\rho(\lambda,\mu) = E[\hat{d}_s(\lambda) - \mu]^2$ for d, a random variable from a $N(\mu,1)$ distribution. The optimal minimax threshold, λ_N^*, can be much smaller than the universal threshold for any particular value of N. The minimax method does a better job at picking up abrupt jumps at the expense of smoothness. Although λ_N^* does not exist in closed form, it can be approximated numerically.

SURE (Stein's Unbiased Risk Estimate) method proposed by Donoho and Johnstone (1995) introduces a scheme that uses the wavelet coefficients at each resolution level j to

choose a threshold λ_j. The explicit form of SURE is as follows:

$$SURE(d, \lambda) = N - 2\sum_{i=1}^{N} I_{[|d_i| \leq \lambda]} + \sum_{i=1}^{N} \min(d_i^2, \lambda^2) \, .$$

Let $d_j = (d_{j,1}, d_{j,2}, \ldots, d_{j,2^j})^T$ be the vector of wavelet coefficients at a resolution level j. Then the *SureShrink* threshold at a resolution level j is given by;

$$\lambda_j^{SURE} = \arg \min_{0 \leq \lambda \leq \lambda^U} SURE(d_j, \lambda)$$

where $\lambda^U = \sqrt{2 \log N}\sigma$ is the universal threshold. Wavelet coefficients smaller than the level-dependent threshold are set to zero. This method is very popular in practice.

These shrinkage methods require an estimate of the standard deviation σ for calculating the threshold value (e.g., $VisuShrink$'s threshold is $(2\ln N)^{1/2}\sigma$). Different estimates of σ will leads to distinct threshold, different number of wavelet coefficients and thus different amount of data reductions. This article uses a robust estimate, $\hat{\sigma} = median(|d_{J,k}| : 1 \leq k \leq N/2)/0.67$ suggested by Donoho and Johnstone (1994), where J is the finest resolution level. Next section proposes two new data reduction methods do not require the estimation of σ.

2.4 Data Reduction Methods - RRE_h and RRE_s

In many engineering applications (e.g., Lada, *et al.* (2002)) of the data de-noising and the AMDL methods, we found that many coefficients were used to achieve a very small signal reconstruction error. By experimenting various numbers of coefficients used in the nonlinear signal approximation methods, we found that many sets of reconstructed signals using a fewer number of coefficients provided a very reasonable approximation to the original data. More importantly, the selected wavelet coefficients were rather representative in most of the data analyses, e.g., chi-square test for process fault detection (e.g., Lada, *et al.* (2002)) or

decision tree analysis for process fault classification. This motivates us to search for a more aggressive "data reduction" method for engineering decision-making applications.

All data de-noising, AMDL and nonlinear signal approximation methods retain the largest M_λ number of coefficients based on some derivations of the threshold λ (e.g., estimated from the noisy data to minimize the expected MSE) or the optimization of an objective function to achieve a higher information value (e.g., AMDL). Our methods will also follow this principle by assuming that large wavelet coefficients (in their absolute value) will better characterize signal patterns and thus retain more information. Our data reduction methods are similar to the AMDL method by selecting M to minimize an objective function balancing the goals of limiting errors in the "signal reconstruction" and more aggressive data reduction.

Definition 1. The energy of a finite sequences $\boldsymbol{f} = (f_1, \cdots, f_N)$ is defined by $\xi = ||\boldsymbol{f}||^2$. Correspondingly, the empirical estimate of the energy of a data signal is $\hat{\xi} = ||\mathbf{y}||^2 = ||\mathbf{d}||^2$.

The following theorem gives an upper bound of the approximation (or estimation) error using the largest M wavelet coefficients (in the absolute values).

Theorem 1. *For $\boldsymbol{f} \in L^2(\mathbb{R})$, an upper bound of the approximation error for $\boldsymbol{f}_M$, is $\|\boldsymbol{f} - \boldsymbol{f}_M\|^2 \leq [(N-M)/M]\ \xi$, and an upper bound of the estimation error for $\hat{\boldsymbol{f}}_M$ is $E\left\|\boldsymbol{y} - \hat{\boldsymbol{f}}_M\right\|^2 \leq [(N-M)/M]\ E(\hat{\xi})$.*

Proof: In this proof, we focus on the stochastic case first, and address the modification of the proof for the deterministic case in the end. Let $d^2_{(1)} \geq d^2_{(2)} \geq \cdots d^2_{(N)}$ be the ordered

energies of wavelet coefficients. Because

$$E(\hat{\xi}) = E\left\|\mathbf{y}\right\|^2 = E\left\|\mathbf{d}\right\|^2 = \sum_{i=1}^{N} E(d_i^2) = \sum_{i=1}^{N} E(d_{(i)}^2) \geq \sum_{i=1}^{M} E(d_{(i)}^2) \geq ME(d_{(M)}^2),$$

the inequalities, $E(d_{(M)}^2) \leq E(\hat{\xi})/M$ holds for $M = 1, 2, \cdots, N$. Therefore,

$$E\left\|\mathbf{y} - \hat{\boldsymbol{f}}_M\right\|^2 = \sum_{i=M+1}^{N} E(d_{(i)}^2) \leq \sum_{i=M+1}^{N} E(\hat{\xi})/i \leq (N-M)E(\hat{\xi})/M.$$

For the deterministic case, replace $d_{(i)}$'s with $\theta_{(i)}$'s, $E(\hat{\xi})$ with $\xi = \left\|\mathbf{f}\right\|^2 = \left\|\boldsymbol{\theta}\right\|^2$, and delete the expectations. The error bound will be derived as stated in Theorem 1.

□

Theorem 2 shows that our methods depend on the energy of the data affecting the data-reduction property, instead of variance (σ^2) of the data noises affecting the data-deniosing properties. Thus, data-reduction and -deniosing methods should be distinct for serving different purposes. Data-denoising procedures aimed to find the estimate $\hat{\theta}$ (and $\hat{\boldsymbol{f}}$) for reducing "modeling error" of $\boldsymbol{f}$. Thus, the data-denoising methods are usually more aggressive in reducing the errors. On the other hand, the data-reduction methods select the "reduced-size" data with a more aggressive data reduction ratio. However, the selected reduced-size data should be representative enough in capturing key data characteristics for subsequent planned or unplanned decision analyses. This motivates the proposal of our data-reduction criteria with goals of balancing two ratios: (1) the relative data-energy capatured in the approximation model, and (2) the relative number of coefficients used, i.e., the data-reduction ratio. Because the first ratio goes down when more coefficients used, our objective function given below will change it to the energy *not* captured for making the

function convex.

$$RRE_h(\lambda) = \frac{\mathrm{E}||\boldsymbol{d} - \hat{\boldsymbol{d}}_h(\lambda)||^2}{\mathrm{E}||\boldsymbol{d}||^2} + \omega \frac{\mathrm{E}||\hat{\boldsymbol{d}}_h(\lambda)||_0}{N}, \tag{9}$$

where $||\hat{\boldsymbol{d}}_h(\lambda)||_0 = \sum_{i=1}^{N} |\hat{d}_{h,i}(\lambda)|_0$ is the number of coefficients selected as the "reduced-size" data, and $|\hat{d}_{h,i}(\lambda)|_0 = 1$, if $\hat{d}_{h,i}(\lambda) \neq 0$; $|\hat{d}_{h,i}(\lambda)|_0 = 0$, otherwise.

The use of "normalizing constants" to make the two balancing terms compatible is critical. The weighting parameter ω is user-selected or provided by method such as generalized cross-validation (GCV) method (Weyrich and Warhola, 1998). However, as experienced from Weyrich and Warhola (1998) further studies are needed for developing the GCV-like selection of ω in our problem and understanding its properties. For simplicity, this article will use $\omega = 1$, which places equal weights in both components in follow-up studies.

The following uses engineering and statistical experience to motivate the objective function. Our discussion will be focused on the *hard-thresholding-based method* RRE_h. A similarly motivated method RRE_s based on the soft-thresholding policy is presented in Appendix. In the wavelet-shrinkage literature, it has been shown that hard-thresholding results in a larger variance of estimates, while soft-thresholding has a larger bias. Hard-thesholding is also very sensitive to small changes in the data. Soft-thresholding has various advantages such as continuity of the shrinkage rule. See Bruce and Gao (1996) for a comparison study between these two thresholding policies in data-denoising applications. See Tables III to V for their comparisons in data-reduction applications.

In engineering applications such as Mallat (1998, pages 378-391), the "relative error,"

$$RE = \frac{||\boldsymbol{f} - \hat{\boldsymbol{f}}||}{||\boldsymbol{f}||}, \quad \text{where} \quad ||\boldsymbol{f}|| = (\sum_{i=1}^{N} f(t_i)^2)^{1/2},$$

is commonly used in comparing signal approximation quality. This article utilizes a thresholding parameter λ to decide which wavelet-domain data to keep and which to discard in decision-making analyses using the terms $\hat{d}_{h,i}(\lambda) = I(|d_i| > \lambda)d_i$, $i = 1, \ldots, N$. Ideally, only a small portion of the data is kept to meet the data-reduction goal. This is quite different to the data-deniosing procedure, where the parameter τ was not set as the threshold originally for the data-reduction purpose in the construction of the objective function (31).

Eq. (9)'s second component serves as a penalty term for limiting the size of data used in follow-up decision analyses. Similar penalty ideas have been used in ridge regression (Hastie *et al.*, 2001, page 59) and neural network (Hastie *et al.*, 2001, page 356). For example, like the data-denoising method of finding estimate $\hat{\theta}$, ridge regression finds the optimal choice of estimate of regression coefficients to minimize the following objective function:

$$\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 + \omega \sum_{j=1}^{p} {\beta_j}^2,$$

where ω is a weighting parameter like the one in Eq. (9). Note that this objective function is not normalized as done in Eq. (9). More importantly, due to the different purposes compared to data-reduction methods, ridge regression does not use a threshold to select which data to keep in follow-up decision analyses.

The following theorems show that our procedures, RRE_h and RRE_s, have hard and soft-thresholding interpretation in the shrinkage methods, respectively, and the thresholding levels depend on signals in terms of their energy. Some asymptotic results are derived. Note that the threshold value of the $VisuShrink$ method is $(2 \ln N)^{1/2}\sigma$, which does not depend on the energy from the signal $\boldsymbol{f}$ or data curve $\mathbf{y}$.

Theorem 2. *Consider the model stated in (6). Then, we have*

(i) the objective function $RRE_h(\lambda)$ is minimized uniquely at $\lambda = \lambda_{N,h}$ where

$$\lambda_{N,h} = (\frac{1}{N}\mathrm{E}||\boldsymbol{d}||^2)^{1/2}; \tag{10}$$

The moment estimate of $\lambda_{N,h}$,

$$\hat{\lambda}_{N,h} = \left(\frac{1}{N}\sum_{i=1}^{N} d_i^2\right)^{1/2} = (\frac{\hat{\xi}}{N})^{\frac{1}{2}}, \tag{11}$$

(ii) $(\hat{\lambda}_{N,h} - \lambda_{N,h}) \xrightarrow{w.p.1} 0$;

(iii) $\sqrt{N}(\hat{\lambda}_{N,h} - \lambda_{N,h})/\sigma^*_{N,h} \xrightarrow{d} N(0,1)$, *where*

$$(\sigma^*_{N,h})^2 = \frac{1}{4N}\left(\frac{4\sigma^2\sum_{i=1}^{N}\theta_i^2 + 2N\sigma^4}{\sum_{i=1}^{N}\theta_i^2 + \sigma^2}\right).$$

Proof: Denote

$$H_i(\lambda) = \mathrm{E}(I(|d_i| \le \lambda)d_i^2) = \int_{-\lambda}^{\lambda} t^2 \frac{1}{\sigma}\phi\left(\frac{t-\theta_i}{\sigma}\right)dt$$

and

$$h_i(\lambda) = \mathrm{E}(|\hat{d}_{h,i}(\lambda)|_0) = \mathrm{E}(I(|d_i| > \lambda) = 1 - \int_{-\lambda}^{\lambda}\frac{1}{\sigma}\phi\left(\frac{t-\theta_i}{\sigma}\right)dt,$$

where $\phi(x) = (2\pi)^{-1/2}\exp(-t^2/2)$, the standard normal density. It follows that

$$\begin{aligned}
\mathrm{E}||\boldsymbol{d} - \hat{\boldsymbol{d}}_h(\lambda||^2) &= \sum_{i=1}^{N}\mathrm{E}(d_i - I(|d_i| > \lambda)d_i)^2 = \sum_{i=1}^{N}\mathrm{E}(I(|d_i| \le \lambda)d_i^2) = \sum_{i=1}^{N} H_i(\lambda), \\
\mathrm{E}||\hat{\boldsymbol{d}}_h(\lambda)||_0 &= \sum_{i=1}^{N}\mathrm{E}(|\hat{d}_i(\lambda)|_0) = \sum_{i=1}^{N}\mathrm{E}(I(|d_i| > \lambda)) = \sum_{i=1}^{N} h_i(\lambda).
\end{aligned}$$

Then, $RRE_h(\lambda)$ can be written as

$$RRE_h(\lambda) = \sum_{i=1}^{N} H_i(\lambda)/\mathrm{E}||\boldsymbol{d}||^2 + \frac{1}{N}\sum_{i=1}^{N} h_i(\lambda).$$

Because of

$$\frac{dh_i(\lambda)}{d\lambda} = -\frac{1}{\sigma}\left[\phi\left(\frac{\lambda-\theta_i}{\sigma}\right) + \phi\left(\frac{-\lambda-\theta_i}{\sigma}\right)\right] < 0$$

and

$$\frac{dH_i(\lambda)}{d\lambda} = \frac{\lambda^2}{\sigma}\left[\phi\left(\frac{\lambda-\theta_i}{\sigma}\right) + \phi\left(\frac{-\lambda-\theta_i}{\sigma}\right)\right] = -\lambda^2\frac{dh_i(\lambda)}{d\lambda},$$

we know that

$$\frac{dRRE_h(\lambda)}{d\lambda} = \left(-\lambda^2/\mathrm{E}(||\boldsymbol{d}||^2) + \frac{1}{N}\right)\sum_{i=1}^{N}\frac{dh_i(\lambda)}{d\lambda} = 0,$$

only if

$$\lambda = \lambda_{N,h} = \left(\frac{1}{N}\mathrm{E}||\boldsymbol{d}||^2\right)^{1/2}.$$

Since d_i's are independently $N(\theta_i, \sigma^2)$ distributed, $N\hat{\lambda}^2_{N,h}/\sigma^2 = \sum_{i=1}^{N} d_i^2/\sigma^2$ is $\chi^2(N, \delta_N)$ distributed with degree of freedom N and non-centrality parameter $\delta_N = \sum_{i=1}^{N}\theta_i^2/\sigma^2$. It follows that $\mathrm{E}(\hat{\lambda}^2_{N,h}) = \sigma^2(\delta_N/N + 1) = \lambda_N$ and $\mathrm{Var}(\hat{\lambda}^2_{N,h}) = \sigma^4(4\delta_N + 2N)/N^2 \to 0$, as $N \to \infty$. Note that $f(t)$ is continuous on $[0, T]$, and then $\max_{0\le t\le T}|f(t)| = K \le \infty$. Because DWT is orthonormal, $|\theta_i|$, $i = 1, 2, \ldots, N$, should be uniformly bounded, as $N \to \infty$. Without loss of generality, we assume that $|\theta_i| < K$, $i = 1, 2, \cdots, N$. Therefore,

$$\lim_{N\to\infty}\sum_{i=1}^{N}\frac{\theta_i^2}{i^2} < K^2\lim_{N\to\infty}\sum_{i=1}^{N}\frac{1}{i^2} < \infty,$$

and we know that

$$\lim_{N\to\infty}\sum_{i=1}^{N}\frac{\mathrm{Var}(d_i^2)}{i^2} < (4\sigma^2K^2 + 2\sigma^4)\lim_{N\to\infty}\sum_{i=1}^{N}\frac{1}{i^2} < \infty.$$

Therefore, from the Kolmogorov Theorem (Serfling, 1980, p.27), we know that $(\hat{\lambda}_{N,h} - \lambda_{N,h}) \xrightarrow{w.p.1} 0$, i.e. the result (ii) is true.

In order to show the asymptotic normality of $\sqrt{N}(\hat{\lambda}^2_{N,h} - \lambda^2_{N,h})/\sigma(\hat{\lambda}^2_{N,h})$, it is sufficient

to verify the following Lindeberg condition (Serfling. 1980, p.30), for every $\varepsilon > 0$

$$\frac{1}{N}\sum_{i=1}^{N}\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt \to 0, \quad N\to\infty, \tag{12}$$

where $\mu_i = \mathrm{E}(d_i^2) = \theta_i^2 + \sigma^2$. It follows that

$$\begin{aligned}
\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt &= O\left(\int_{t^2>\varepsilon\sqrt{N}} t^4\phi\left(\frac{t-\theta_i}{\sigma}\right)dt\right)\\
&= O\left(\int_{t>\varepsilon^{1/2}N^{1/4}} t^4\phi\left(\frac{t-\theta_i}{\sigma}\right)dt\right)\\
&= O\left(\varepsilon^2 N\phi\left(\frac{\varepsilon^{1/2}N^{1/4}-\theta_i}{\sigma}\right)\right)\\
&= O\left((\varepsilon^2 N\exp\left\{-\frac{\varepsilon\sqrt{N}}{2\sigma^2}\right\}\right).
\end{aligned}$$

Therefore, for every $\varepsilon > 0$, as $N\to\infty$,

$$\frac{1}{N}\sum_{i=1}^{N}\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt = O\left((\varepsilon^2 N\exp\left\{-\frac{\varepsilon\sqrt{N}}{2\sigma^2}\right\}\right)\to 0,$$

and we know that $\sqrt{N}(\hat{\lambda}_{N,h}^2 - \lambda_{N,h}^2)/\sigma(\hat{\lambda}_{N,h}^2)$ is asymptotically normal. Then, from the delta method, if $(T_N-\eta_N)/\tau_N \xrightarrow{d} N(0,1)$, then $[h(T_N)-h(\eta_N)]/[\tau_N h'(\eta_N)] \xrightarrow{d} N(0,1)$ provided h is continuous function such that $h'(\eta_N)$ exists and $h'(\eta_N)\neq 0$. In our situation, let $T_N = \hat{\lambda}_{N,h}^2$, $\eta_N = \lambda_{N,h}^2$, and $\tau_N = \sigma_N(\hat{\lambda}_N^2)$, $h(\eta) = \sqrt{\eta}$ and $h'(\eta) = 1/2\sqrt{\eta}$, by applying the delta method, we can get the stated results of (iii).

□

Similar idea presented for RRE_h can be extended from the soft-thresholding idea. Its analytical properties can be derived similarly as presented in Theorem 3. Denote by $\hat{\boldsymbol{d}}_s(\lambda) = (\hat{d}_{s,1}(\lambda),\ldots,\hat{d}_{s,N}(\lambda))^\top$, where $\hat{d}_{s,i}(\lambda) = I(|d_i|>\lambda)sign(d_i)(|d_i|-\lambda)$, $i=1,\ldots,N$. Then,

$$RRE_s(\lambda) = \frac{\mathrm{E}||\boldsymbol{d}-\hat{\boldsymbol{d}}_s(\lambda)||^2}{(\mathrm{E}||\boldsymbol{d}||^2)^{\frac{1}{2}}} + \omega\frac{\mathrm{E}||\hat{\boldsymbol{d}}_s(\lambda)||_1}{(\mathrm{E}||\boldsymbol{d}||_1)^{\frac{1}{2}}}, \tag{13}$$

where $||\hat{\boldsymbol{d}}_s(\lambda)||_1 = \sum_{i=1}^{N} |\hat{d}_{s,i}(\lambda)|$.

Theorem 3. *Consider the model stated in (26). Then, we have*

(i) the objective function $RRE_s(\lambda)$ is minimized uniquely at $\lambda = \lambda_{N,s}$ where

$$\lambda_{N,s} = 0.5 * (\frac{\mathrm{E}||\boldsymbol{d}||^2}{\mathrm{E}||\boldsymbol{d}||_1})^{1/2}; \tag{14}$$

The empirical estimate of $\lambda_{N,s}$,

$$\hat{\lambda}_{N,s} = 0.5 * \left(\frac{\sum_{i=1}^{N} d_i^2}{\sum_{i=1}^{N} |d_i|} \right)^{1/2} = 0.5 * (\frac{\hat{\xi}}{l_1})^{\frac{1}{2}}, \tag{15}$$

where l_1 is the L_1-norm of $\boldsymbol{d}$.

(ii) $(\hat{\lambda}_{N,s} - \lambda_{N,s}) \xrightarrow{w.p.1} 0$;

Proof: Denote

$$V_i(\lambda) = \mathrm{E}(|\hat{d}_{s,i}(\lambda)|) = \mathrm{E}\left((|I(|d_i| > \lambda) sign(d_i)(|d_i| - \lambda)|\right).$$

According to the intervals of d_i, the term $I(|d_i| > \lambda) sign(d_i)(|d_i| - \lambda)$ can be defined as follows:

$$I(|d_i| > \lambda) sign(d_i)(|d_i| - \lambda) = \begin{cases} d_i + \lambda, & d_i < -\lambda \\ 0, & -\lambda < d_i < \lambda \\ d_i - \lambda, & d_i > \lambda. \end{cases}$$

Then,

$$\begin{aligned} V_i(\lambda) &= \mathrm{E}(|I(d_i > \lambda)(d_i - \lambda)|) + \mathrm{E}(|I(d_i < -\lambda)(d_i + \lambda)|) \\ &= \int_{\lambda}^{\infty} \frac{|t - \lambda|}{\sigma} \phi\left(\frac{t - \theta_i}{\sigma}\right) dt + \int_{-\infty}^{-\lambda} \frac{|t + \lambda|}{\sigma} \phi\left(\frac{t - \theta_i}{\sigma}\right) dt. \end{aligned}$$

Since,

$$\begin{aligned}
\mathrm{E}(d_i - \hat{d}_{s,i}(\lambda))^2 &= \mathrm{E}\left[(d_i - I(|d_i| > \lambda) sign(d_i)(|d_i| - \lambda))^2\right] \\
&= \mathrm{E}\left[I(|d_i| \leq \lambda) d_i^2\right] + \lambda^2 \mathrm{E}\left[I(|d_i| > \lambda)\right] \\
&= H_i(\lambda) + \lambda^2 h_i(\lambda),
\end{aligned}$$

$RRE_s(\lambda)$ can be written as

$$RRE_s(\lambda) = \left(\sum_{i=1}^{N} H_i(\lambda) + \lambda^2 \sum_{i=1}^{N} h_i(\lambda)\right) / \mathrm{E}(||\boldsymbol{d}||^2)^{\frac{1}{2}} + \sum_{i=1}^{N} V_i(\lambda) / \mathrm{E}(||\boldsymbol{d}||_1)^{\frac{1}{2}}.$$

Since

$$\begin{aligned}
\frac{dV_i(\lambda)}{d\lambda} &= -\frac{\lambda}{\sigma}\phi\left(\frac{\lambda - \theta_i}{\sigma}\right) - \int_{\lambda}^{\infty} \frac{1}{\sigma}\phi\left(\frac{t - \theta_i}{\sigma}\right) dt + \frac{\lambda}{\sigma}\phi\left(\frac{\lambda - \theta_i}{\sigma}\right) \\
&\quad + \frac{\lambda}{\sigma}\phi\left(\frac{-\lambda - \theta_i}{\sigma}\right) - \int_{-\infty}^{-\lambda} \frac{1}{\sigma}\phi\left(\frac{t - \theta_i}{\sigma}\right) dt - \frac{\lambda}{\sigma}\phi\left(\frac{-\lambda - \theta_i}{\sigma}\right) \\
&= -\mathrm{E}(|d_i| > \lambda) = -h_i(\lambda),
\end{aligned}$$

$$\begin{aligned}
\frac{dRRE_s(\lambda)}{d\lambda} &= \left[-\lambda^2 \sum_{i=1}^{N} \frac{dh_i(\lambda)}{d\lambda} + 2\lambda \sum_{i=1}^{N} h_i(\lambda) + \lambda^2 \sum_{i=1}^{N} \frac{dh_i(\lambda)}{d\lambda}\right] / \mathrm{E}(||\boldsymbol{d}||^2)^{\frac{1}{2}} \\
&\quad - \left[\sum_{i=1}^{N} h_i(\lambda)\right] / \mathrm{E}(||\boldsymbol{d}||_1)^{\frac{1}{2}} \\
&= \left(\frac{2\lambda}{\mathrm{E}(||\boldsymbol{d}||^2)^{\frac{1}{2}}} - \frac{1}{\mathrm{E}(||\boldsymbol{d}^{|}|_1)^{\frac{1}{2}}}\right) \sum_{i=1}^{N} h_i(\lambda) = 0,
\end{aligned}$$

only if

$$\lambda = \lambda_{N,s} = \frac{1}{2}\left(\frac{\mathrm{E}||\boldsymbol{d}||^2}{\mathrm{E}||\boldsymbol{d}^{|}|_1}\right)^{1/2}.$$

Table 1: Results of Data Reduction for Testing Signals

Signals	Energy	Threshold value		M				
		$\hat{\lambda}_h$	$\hat{\lambda}_s$	$RREh$	$RREs$	$Visu$	$Risk$	$SURE$
Nason	94.25	0.3034	0.6986	31	138	192	225	324
Heavisine	90.28	0.2969	0.6803	28	143	287	290	292
Blocks	72.36	0.2658	0.5099	67	379	389	407	518
Bumps	17.63	0.1312	0.3401	91	405	646	664	722

Also, similar to the proof of (ii) of Theorem 2, we know that $(\hat{\lambda}_{N,s} - \lambda_{N,s}) \xrightarrow{w.p.1} 0$ from the Kolmogorov Theorem and Slutsky's Theorem, i.e. the result (ii) is true.

□

Consider a few well-known testing signal curves which are "normalized" forms (in the same scale and zero mean)) taken from the literature (e.g., Donoho and Johnstone, 1995). Table 1 shows the relationship between the energy value of signal and data reduction. In general, if the signal has larger value of energy, its threshold value will be higher (see the threshold values for RRE_h and RRE_s for examples), then it has large chance to have a smaller M. However, because the threshold values for the RRE_h in Nason and Heavisine signals are very close, when the following "unbalancing" factor comes in, we do see some exceptions for the Heavisine signal, e.g., slightly smaller M for the RRE_h. If most of the signal energy is kept in a few larger wavelet coefficients (with relatively many small-coefficients) then the signal has a set of very "unbalanced" wavelet coefficients. When there are more number of smaller coefficients, the number of thresholded coefficients is smaller. This leads to a smaller M. See Vidakovic (2000) for a technique to compare signals with different unbalancing characteristics.

Table 2 presents the impact of not using the normalizing constants in the proposed

Table 2: Impacts of Normalization for Data Reduction

Signals	With Normalization			Without Normalization		
	Relative error	M/N	RRE_h	Relative error	M/N	RRE_h
Bumps ($SNR^* = \infty$)	2.18E-02	0.090	0.112	2.81E-19	0.770	0.770
Bumps ($SNR^* = 15$)	2.94E-02	0.066	0.096	6.18E-04	0.456	0.456
Bumps ($SNR^* = 7$)	3.97E-02	0.066	0.106	2.98E-03	0.432	0.435
Bumps ($SNR^* = 3$)	9.45E-02	0.066	0.161	1.60E-02	0.395	0.411
RTCVD	1.77E-02	0.130	0.147	8.89E-07	0.578	0.578
Antenna	4.25E-02	0.180	0.222	3.27E-05	0.644	0.644

objective function (5), denoted as $RRE_h{}^*$. Without the normalization, the data-reduction ratios are very poor for all cases studied including two real-life data sets from the RTCVD and antenna manufacturing processes, and noisy data simulated from bumps signals, where the notation SNR^* in Table 2 represents the noise level of data. Smaller SNR^* means that the data is noisier. Their behaviors are similar to the use of data-denoising methods for data-reduction purpose. Their relative errors are very small with plots produced by data-denoising methods. On the other hand, although the relative errors in the RRE_h method is larger, all the reconstructed curves capture the main data patterns such as the 11 bumps in Figure 3 and the cups and lobes of the antenna data in Figure 4. This demonstrates that the normalizing constants have a great impact to the data-reduction. See next section for more details of comparisons between data-reduction and -denoising methods.

2.5 Comparisons of the Data Reduction Methods

All of the above six methods could be used for the data reduction purpose. This section evaluates their effectiveness with a few well-known testing signals shown in Figure 2 and two real-life data curves shown in Figures 9 and 10. These signals characterize different

important features of the inhomogeneous signals arising in imaging, seismography, manufacturing and other engineering fields. Symmlet-8 is used in wavelet transforms for all cases. Tables 3-4 present the comparison results with the following summary measures: (1) Reduction ratio (%) : $RR = (1 - M/N) \times 100$; (2) $RelErr = \|\boldsymbol{f} - \hat{\boldsymbol{f}}_M\|/\|\boldsymbol{f}\|$ for the case without random errors and $RelErr = \|\boldsymbol{y} - \hat{\boldsymbol{f}}_M\|/\|\boldsymbol{y}\|$ for the case with random errors; and (3) $AMDL$ quantity.

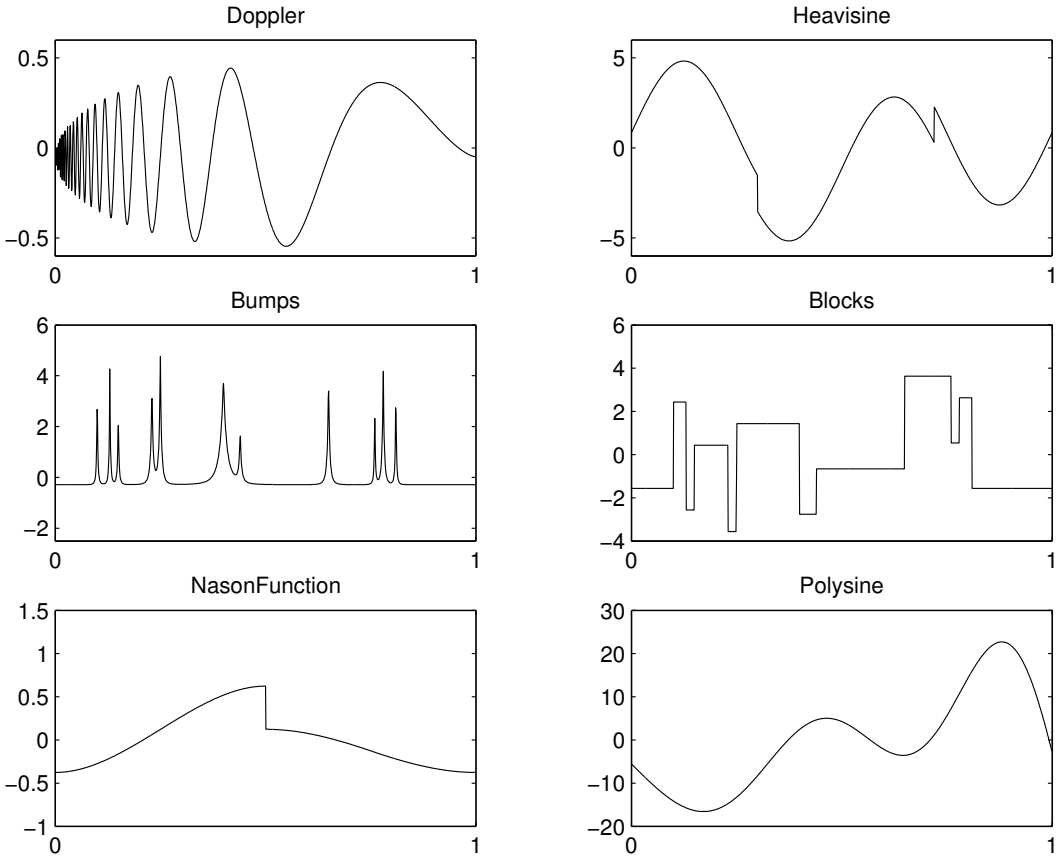

Figure 2: Six Testing Signals from the Literature

Figure 3 shows the results of these data reduction methods applied to the bumps signal. Excellence in limiting the modeling errors is shown in $VisuShrink$, $RiskShrink$, $SURE$ and $AMDL$. RRE_s did as well as the others. RRE_h missed some details in the smoother signal between peaks. However, all the shapes and locations of the 11 peaks were identified

and modeled well by the more aggressive RRE_h method, which has a 90% data reduction ratio as opposed to the 60% in RRE_s and below 40% in all other methods (see Table 3 for details of errors and ratios). Similar conclusions were observed for many other testing signals (see Figure 4 - Figure 8). Case studies show that RRE_h and RRE_s methods did give accurate decision results even with an aggressive data reduction emphasis. The following checks if the proposed methods work well in the two real-life data sets, where errors were involved.

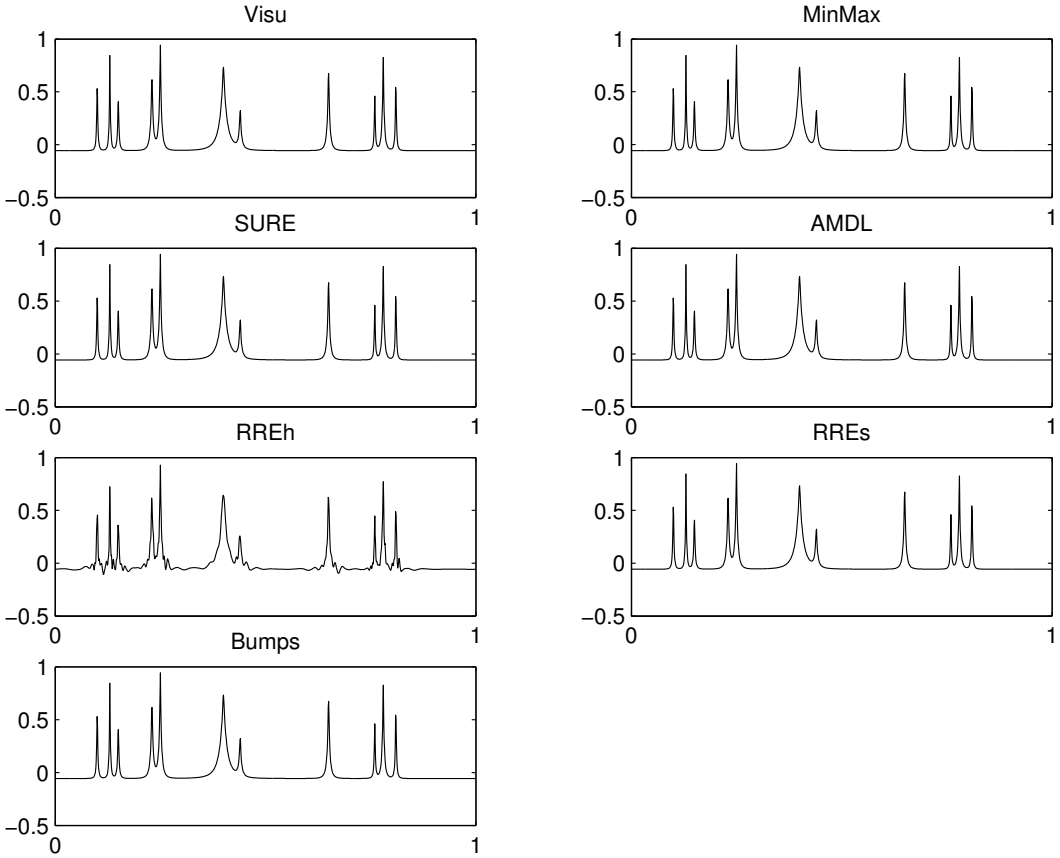

Figure 3: Reconstruction of the Noisy-free Bumps Signal

Example (RTCVD Data). The RTCVD process deposits thin films on the wafer by a temperature driven surface chemical reaction. As feature size decreases, functional operation of semiconductors (e.g., transistors) becomes increasingly susceptible due to variations of

Table 3: Results for the Bumps Signal

Method	M	$RelErr$	RR	$ADML$
$VisuShrink$	646	$1.50E-16$	36%	16390.6
$RiskShrink$	664	$1.23E-18$	35%	13108.3
$SureShrink$	722	$2.22E-21$	29%	26321.8
$AMDL$	894	$3.91E-25$	13%	5506.6
RRE_h	91	$2.18E-02$	91%	32151.2
RRE_s	405	$1.51E-09$	60%	24682.6

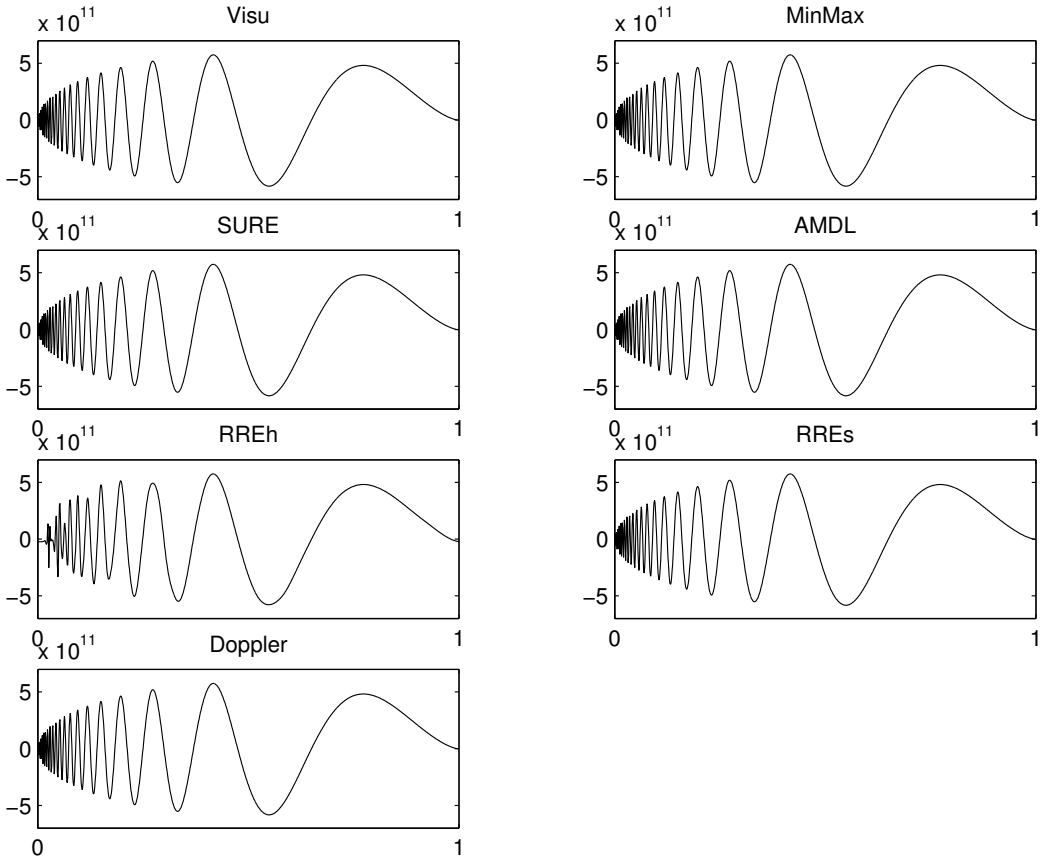

Figure 4: Reconstruction of the Doppler Signal .

deposition processes. Thus, controlling the processing variability is critical. QMS is commonly used in semiconductor manufacturing processes for monitoring thin-film deposition quality. The data shown in Figure 9 is one of the several nominal RTCVD process runs in a research project (Rying, 2001) of developing an "in-situ" measurement technique for online process monitoring. Although there are only 128 data points in the curve, and the

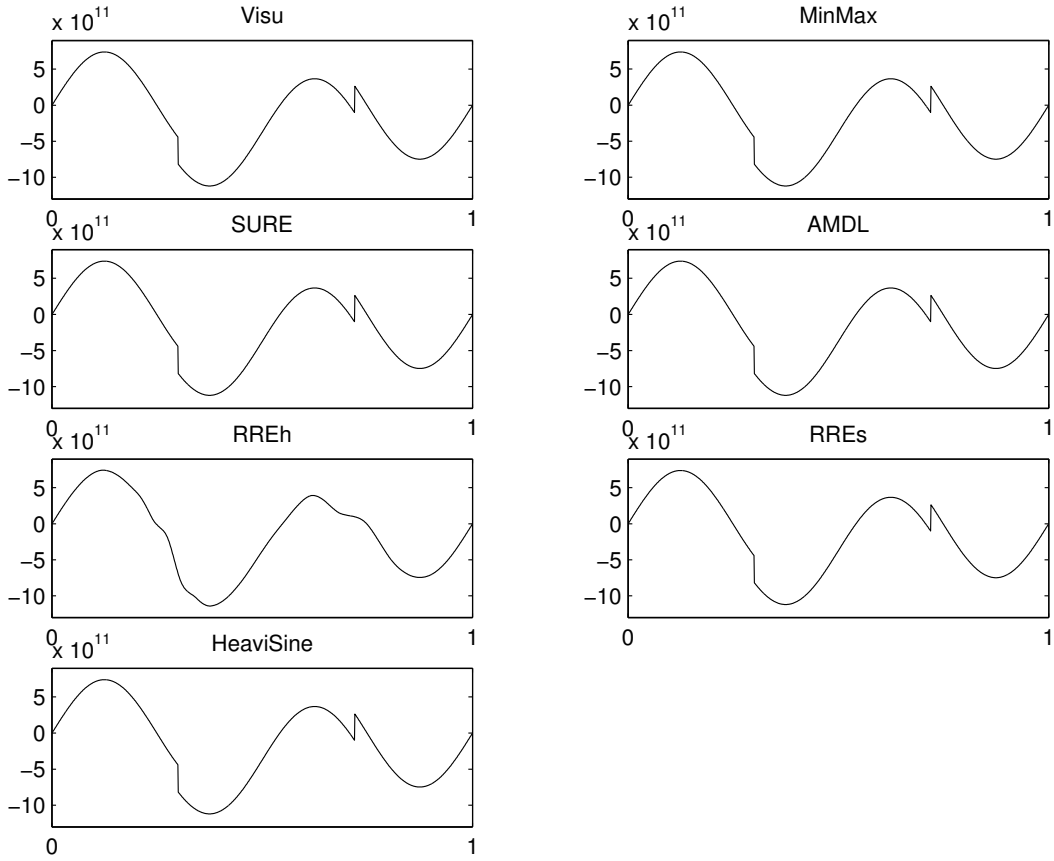

Figure 5: Reconstruction of the Heavisine Signal .

data change-pattern is not very complicated, this case study serves as a basis for developing process monitoring and fault detection/classification tools applicable in many engineering applications. More importantly, wavelet transforms are proven to be useful in locating those change-points, e.g., the two peaks, for developing an integrated metric essential for the in-situ measurement tool. See Rying (2001) for details.

Results in Figure 9 and Table 4 show that RRE_h could be too aggressive in data reduction (87% ratio) due to its non-smoothing fit in the straight rising component (data between 20 to 30 points). However, it did roughly pick up the two peaks and other change-points. AMDL did a much better job in balancing the data reduction ratio and the modeling error in this case. The errors of the three shrinkage methods are smaller (paid a price of a lower

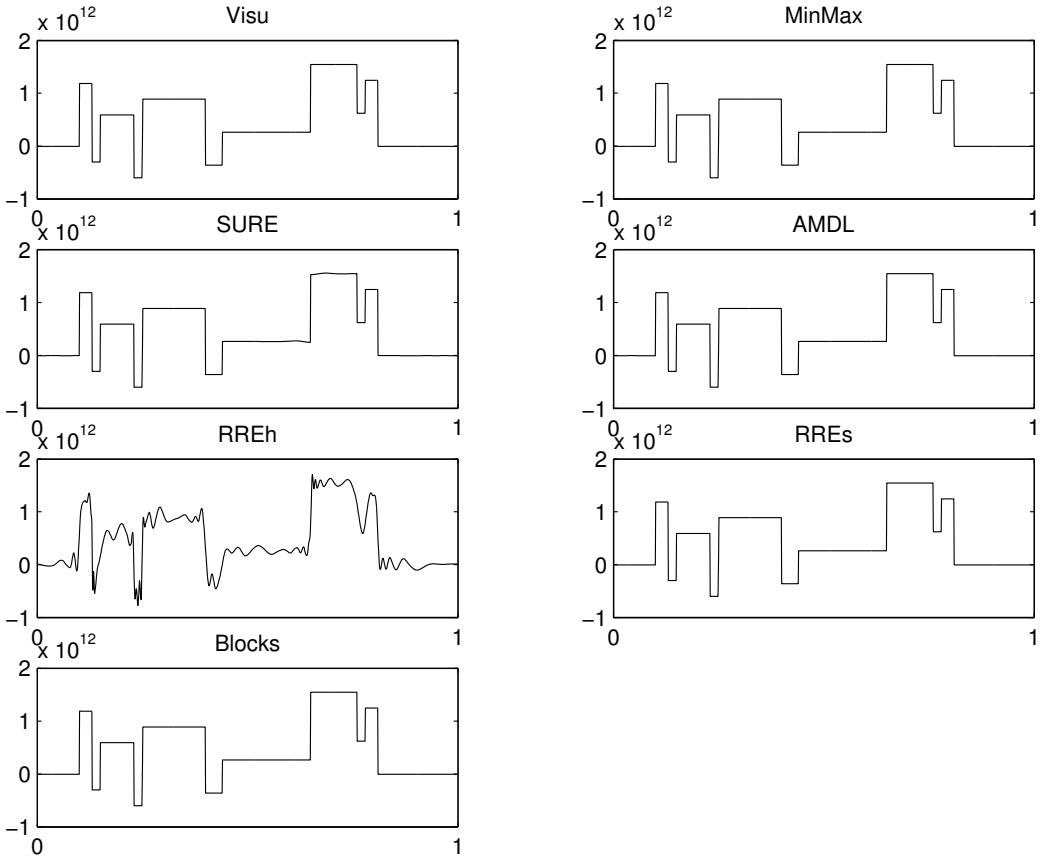

Figure 6: Reconstruction of the Blocks Signal .

reduction ratio). However, inspecting Figure 9, it is difficult to distinguish these errors against AMDL and RRE_s with human eyes.

Example (Antenna Data). With the increasing popularity of wireless communications, a high degree of quality for antenna equipment is needed. Many sets of antenna data like what has been shown in Figure 1 were collected at Nortel for developing a procedure to monitor antenna manufacturing quality. Figure 10 shows the reconstructed antenna curves based on various data reduction methods. Excluding the RRE_h method, all methods model the complicated peak and valley patterns very well. RRE_h provides a reasonable fitting other than the valleys between the second and the third peaks from the main lobe in the

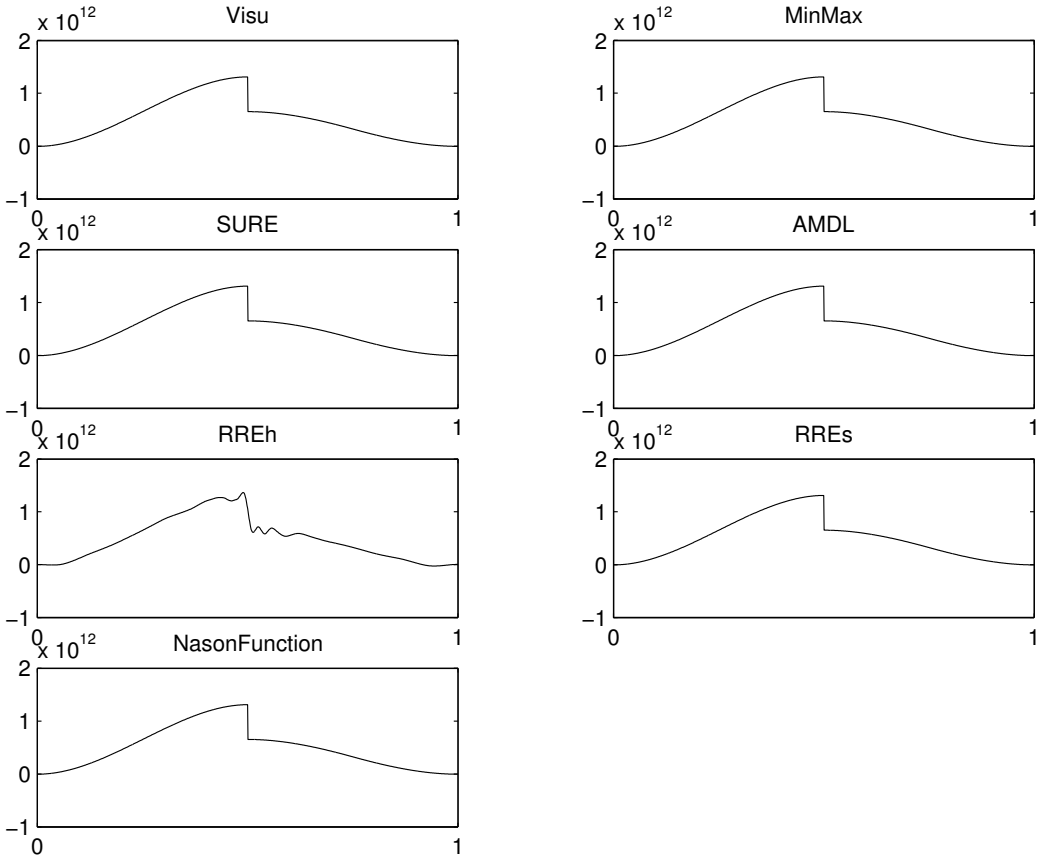

Figure 7: Reconstruction of the Nason's Function .

middle. Surprisingly, the AMDL has an excellent reduction ratio (81%) as good as RRE_h.

Remarks and Discussions:

1. We also test the robustness of the above data-reduction methods against random noises. In a series of experiments, various amount of random normal noises were added to the testing signals. Define SNR^* as $std(\boldsymbol{f})/\sigma$, where $std(\boldsymbol{f})$ is the standard deviation of the discretized signal points, and σ is the standard deviation of noise. Figure 11 shows the noisy bumps with different values of SNR^*'s. Table 5 summarizes model fitting and data-reduction results from all methods in the cases of $SNR^* = 3$, $SNR^* = 7$, and $SNR^* = 15$. Smaller SNR^* means a noisier signal. For the signals with larger SNR^* (less noisy), the

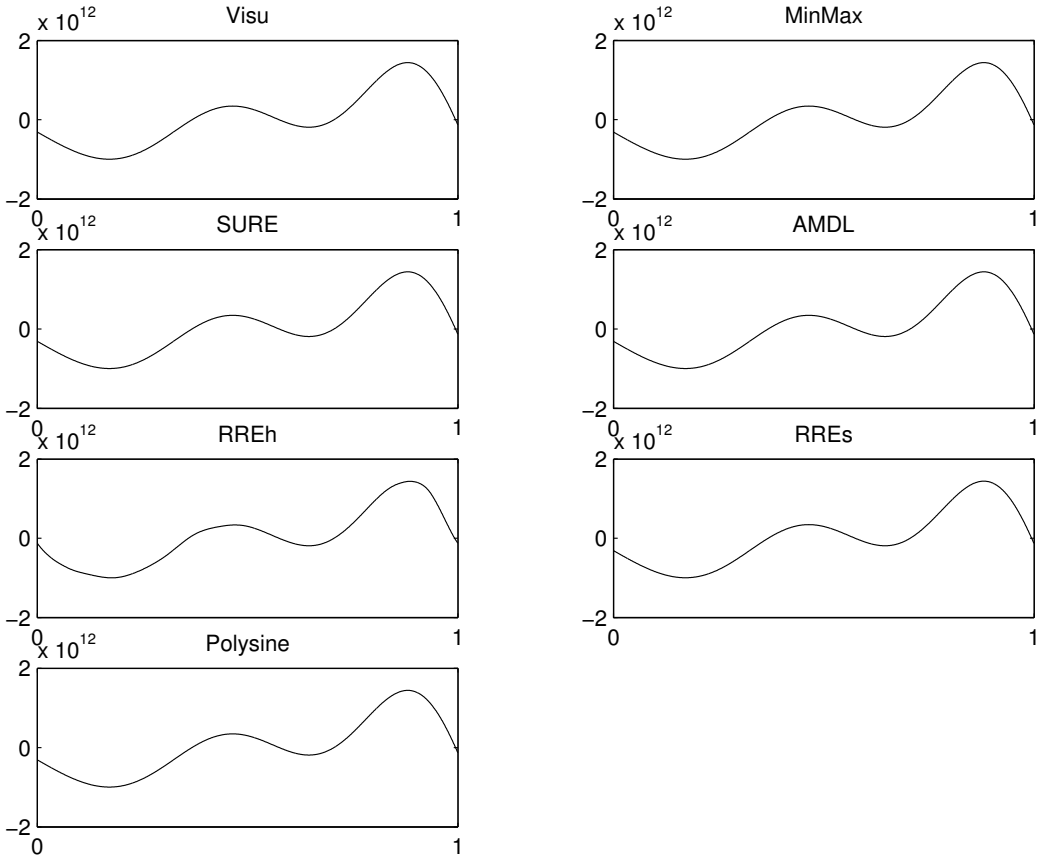

Figure 8: Reconstruction of the Polysine Signal .

noise level (σ) is lower and the threshold value should be lower (e.g., the threshold value of $VisuShrink$ is $2\ln N)^{\frac{1}{2}}\sigma$). This leads to a larger number of selected coefficients. For this reason, the denoising methods are less effective in data-reduction, and use a larger number of wavelet coefficients in the model. See the drops of data-reduction ratio for $SureShrink$ in Table 5 from $SNR^* = 3$ to $SNR^* = \infty$ cases for a specific example. With noisy data, the difference in modeling errors from these six methods is smaller than the difference in the case without added noises, where SNR^* is equal to ∞. The reduction ratio stays the same for the RRE_h, but improved considerably for all other methods. However, they pay a price to have much larger modeling errors (see Table 5) as compared with the results given in Table 3. Surprisingly, the modeling errors from $VisuSbrink$ and $AMDL$ methods in the

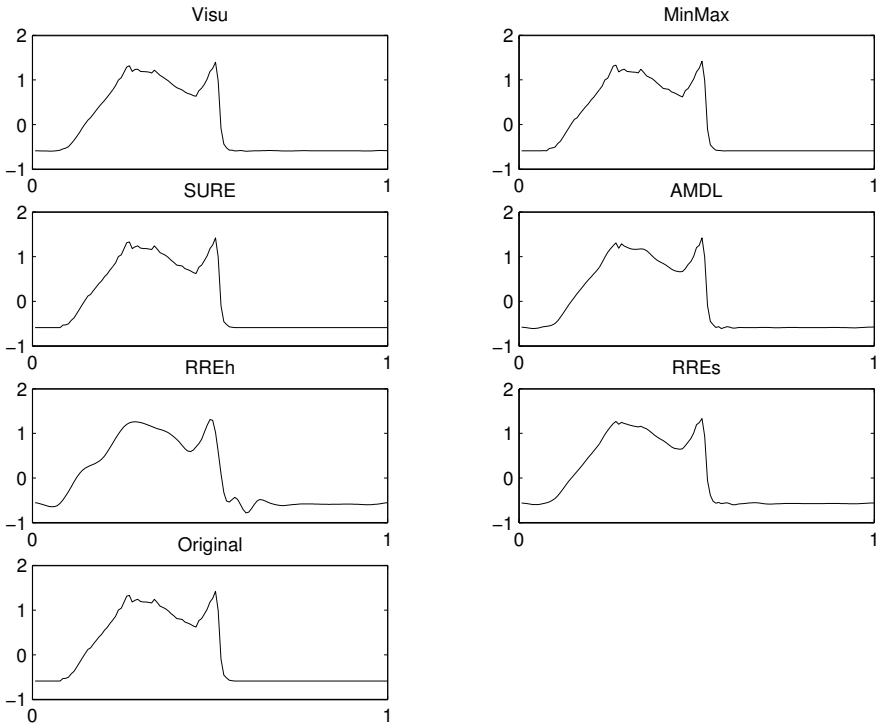

Figure 9: Reconstruction of the RTCVD Signal

case for $SNR^* = 3$ (the most noisy case studied) are larger than the errors in the proposed RREh and RREs methods.

2. In engineering applications such as Lada *et al.* (2002), replicated signal curves exhibit patterns as shown in Figure 12 (a) from the RTCVD experiment. This type of "curve-replicates" could shift "up," "down," "left" or "right" (with some minor pattern changes), but the "overall characteristics" remain similar. This could be easily experienced from the

Table 4: Results for the RTCVD and Antenna Data

	RTCVD		Antenna	
Method	RR	$RelErr$	RR	$RelErr$
$VisuShrink$	50%	$9.92E-05$	59%	$1.70E-03$
$RiskShrink$	46%	$2.37E-06$	45%	$1.07E-04$
$SureShrink$	36%	$8.69E-08$	27%	$1.46E-05$
$AMDL$	75%	$5.35E-04$	81%	$7.47E-03$
RRE_h	87%	$1.77E-02$	82%	$4.25E-02$
RRE_s	68%	$2.27E-03$	67%	$5.55E-03$

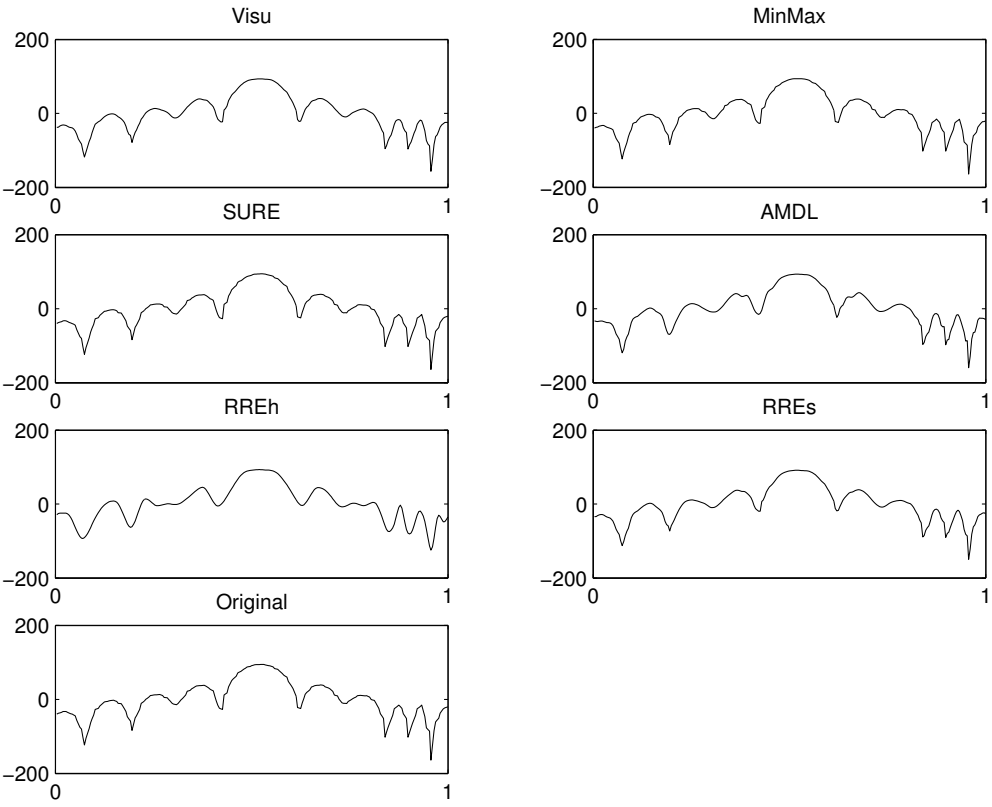

Figure 10: Reconstruction of the Antenna Data

example that the x-ray image of a product is a circle signal. With certain amount of process noises, the resulted circles could have different sizes of radius and distinct centers. But, they are all similar circles. This type of replicates is quite different to replicates generated from the traditional model, where random noises are added to a deterministic functional curve. Figure 12 (b) presents this type of curve with random $N(0, 0.01)$ noises added to a typical RCTVD data. Thus, in the decision tree evaluation experiment, the replicates will be generated based on "engineering noises" using various amount of shifting to simulate "curve-replicates." Then, statistical normal random noises are added. Figures 12 (c) and (d) show one example of the original and the replicated curve.

3. In deciding which wavelet family is most suitable for representing a data signal, the more "disbalancing" type (more separation in the larger and smaller wavelet coefficients) of

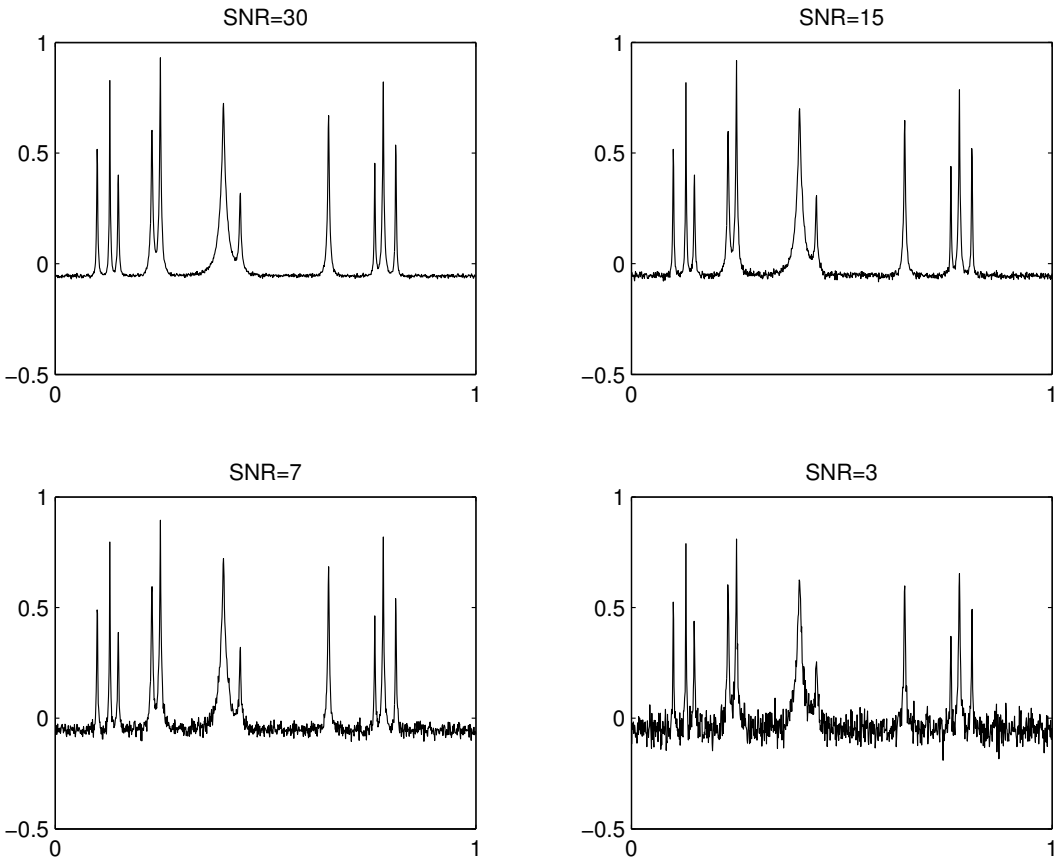

Figure 11: Noisy Bumps Signal at Various Noise Levels

wavelet family is used, the more efficient the data-reduction will be. Because "Symmlet-8" showed excellent disbalancing properties on most of the curves studied in our evaluation studies and application examples, we used it as the "default" choice of the wavelet family in our data-reduction exercises.

In summary, RRE_h, $AMDL$ and RRE_s are more suitable for data reduction purposes. However, RRE_h could be too aggressive in some cases where certain details are ignored; $AMDL$ is not suitable for signal curves "without noise". $VisuShrink$, $RiskShrink$ and $SureShrink$ are not very effective in data reduction, but their modeling quality is excellent. When random normal noises are added to the deterministic signal curves, the difference between these six methods in their modeling quality and data reduction ratio becomes

Table 5: Results for the Noisy Bumps Signal

	$SNR^* = \infty$		$SNR^* = 15$		$SNR^* = 7$		$SNR^* = 3$	
Method	RR	$RelErr$	RR	$RelErr$	RR	$RelErr$	RR	$RelErr$
$Visu$	36%	1.50E-16	85%	1.12E-02	88%	4.18E-02	91%	1.54E-01
$Risk$	35%	1.23E-18	78%	2.52E-03	83%	1.21E-02	86%	6.24E-02
$SURE$	29%	2.22E-21	54%	8.00E-04	70%	8.42E-03	78%	4.91E-02
$AMDL$	13%	3.91E-25	87%	6.37E-03	90%	2.39E-02	95%	1.36E-01
RRE_h	91%	2.18E-02	93%	3.00E-02	93%	4.00E-02	93%	9.45E-02
RRE_s	60%	1.51E-09	85%	1.17E-02	88%	3.94E-02	76%	7.63E-02

smaller, especially when the data are noisier with a smaller SNR. The next section further examines the effectiveness of the data reduction methods with various decision rules.

2.6 Signal Classification Using Reduced-size Data

This section presents the examples of using the selected wavelet coefficients as the reduced-size data for detecting process faults and classifying process fault types. These activities are important in many engineering applications. In particular, when manufacturing processes or systems become very complicated, human operators have difficulty in identifying sources creating process problems. Effective use of process data (e.g., control signals and various stages of process performance measurements) in a timely manner could drastically reduce process defects, production costs or more serious process problems. This section shows the possibility of making excellent decisions on process fault types with classification and regression tree (CART).

CART is a tree with many nodes at various levels in a hierarchy making binary decisions based on values of a chosen variable at each node. CART is very popular in many data mining applications, e.g., customer relationship management. See Breiman *et al.* (1984) for details of CART tree-building and pruning procedures.

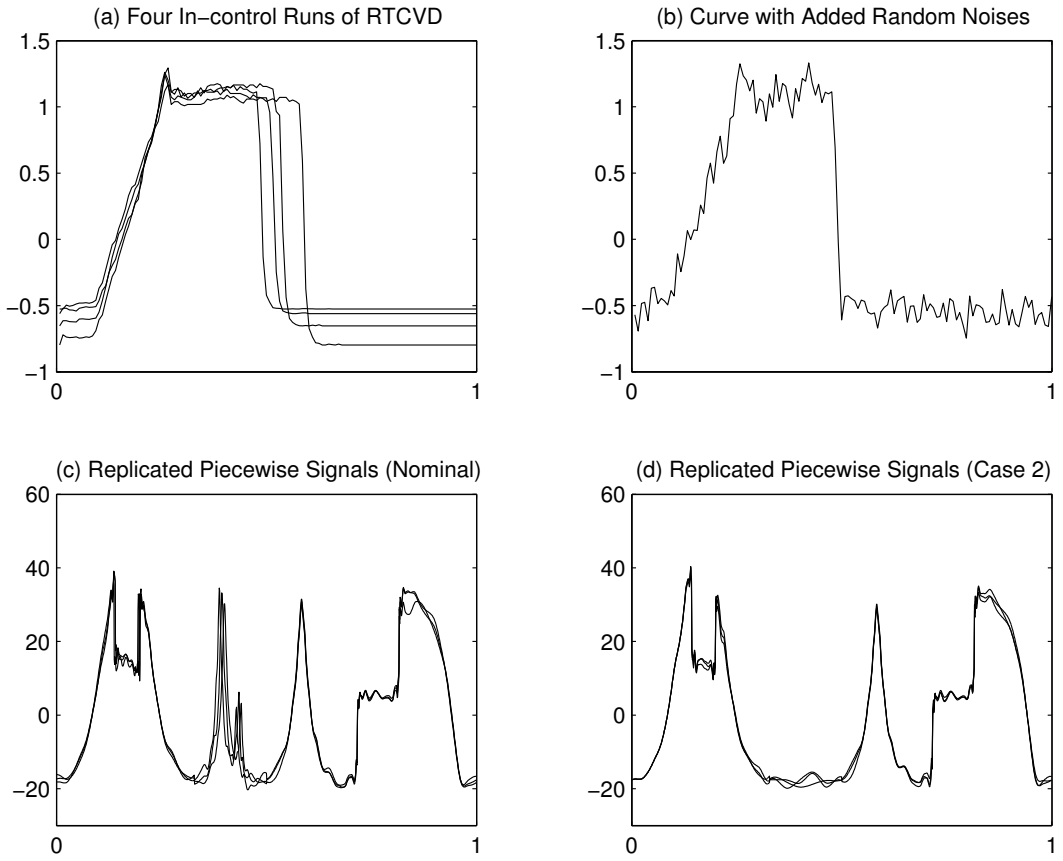

Figure 12: Different Types of Signal Replications

To evaluate the error rate in applying CART to the reduced-size data for classifying process fault types, various "curve-replicates" were generated from a very difficult signal pattern taken from Mallat (1998; page 378). In our experiment, the entire curve is shifted to the left (or right) in 5 (or 10, 15, 20, 25, 30) time-units (out of a total of $N = 1,024$ units) for generating a new curve with added random normal(0, σ^2) noises with $\sigma = 0.1$. See Figures 12 (c) and (d) for one example of the original and replicated data curves.

Figure 13 presents seven fault classes of curves. Some of them are considerably more difficult than the others for decision trees to correctly identify its fault classes. For example, the only difference between fault class 4 and the original curve is a smaller amount of vertical drop of the first rectangle-shape dip around 147 to 204 time units. Class 1 could also be considered a difficult case where the first dip is filled smoothly. For all curves in these eight cases, the data replication method presented in Remark **3** was applied to generate 2,400 total replicated-curves (300 in each case). For dealing with multiple classes of replicated data curves, our study uses the union positions of all thresholded coefficients (obtained from application of RRE methods to individual data curves) to create the reduced-size data. Then, CART is supposed to identify all these fault types successfully based on the reduced-size data obtained from the RRE_s method with a 91.89% reduction ratio.

There is no good guideline available on how to divide the available 2,400 samples into training and testing data sets. Fukunaga (1990) provided arguments in favor of using more samples for testing than for training the classifier. Thus, our experiment used 1/3 of the data randomly selected from each case for training and 2/3 data for testing. Figure 14 shows the CART tree constructed in the wavelet domain using the training data set. This tree has eight terminal nodes, and only four among 83 wavelet coefficients were used for

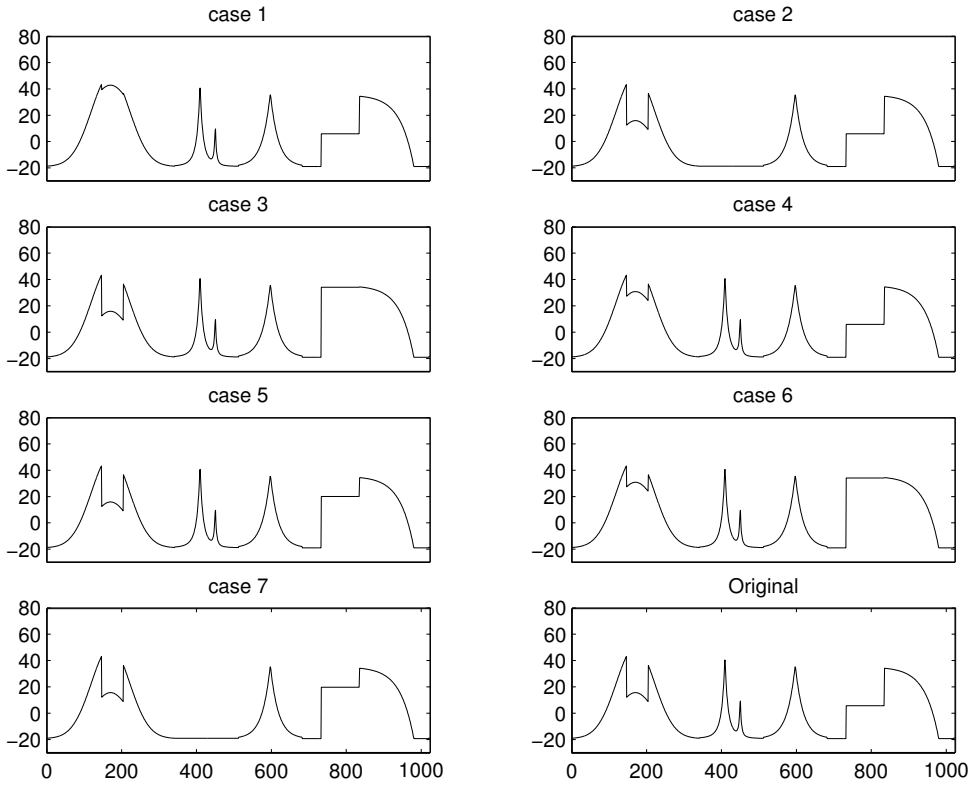

Figure 13: Mallat's Piecewise Signals

classification in this experiment.

In Figure 14 , the notation $c_{L,k}$ means the wavelet coefficients in the coarser level and $d_{j,k}$ the finer wavelet coefficients at the resolution level j and kth position. The first split is $c_{5,6} \leq -28.967$ where $c_{5,6}$ is the $6th$ position coefficient in the coarser resolution level. This coefficient covers the support $[161, 192]$ in the time domain, which is somewhere close to the first rectangle-dip discussed above. In node 2, if $c_{5,17} \leq 52.95$, then the signal is classified into class 2; otherwise, the signal is classified into class 7. The coefficient $c_{5,17}$ covers the support $[513, 544]$, which is slightly to the right of the middle of the curve. This decision rule seems to pick up the major difference in the first dip and the second/third missing peaks (in the middle) from these two classes. Similar interpretation could be obtained for other nodes. In practice, the majority of patterns could be identified in the coarser resolution level,

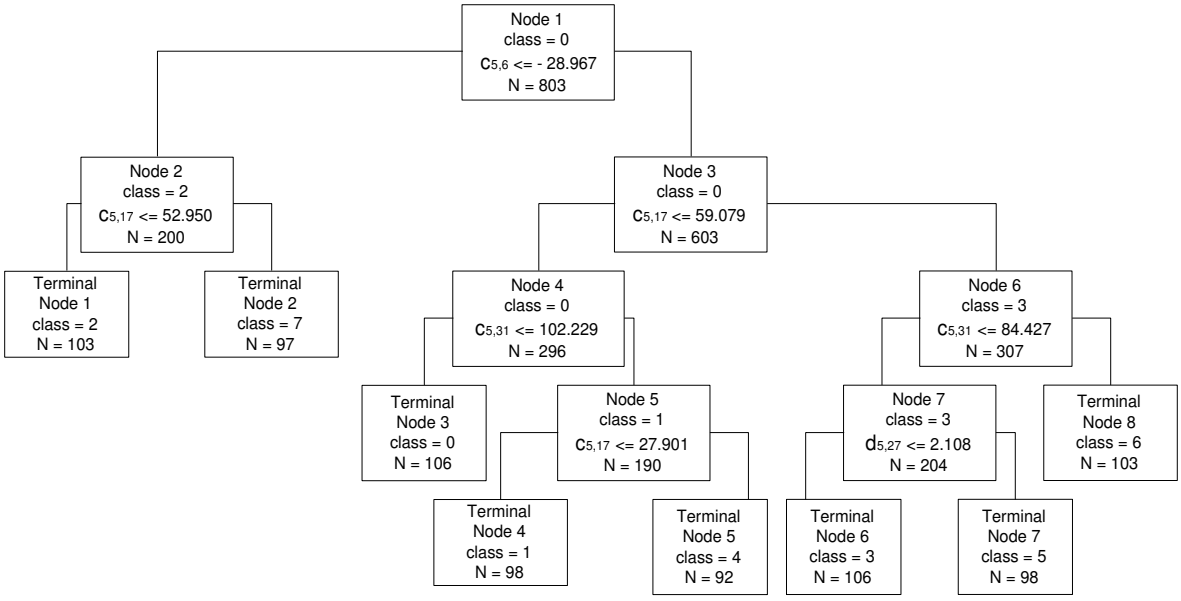

Figure 14: CART Tree in the Wavelet Domain

while only few patterns will require information at finer levels for decision. This provides a hierarchical multi-resolution decision making opportunity not available in the time domain based on the original data.

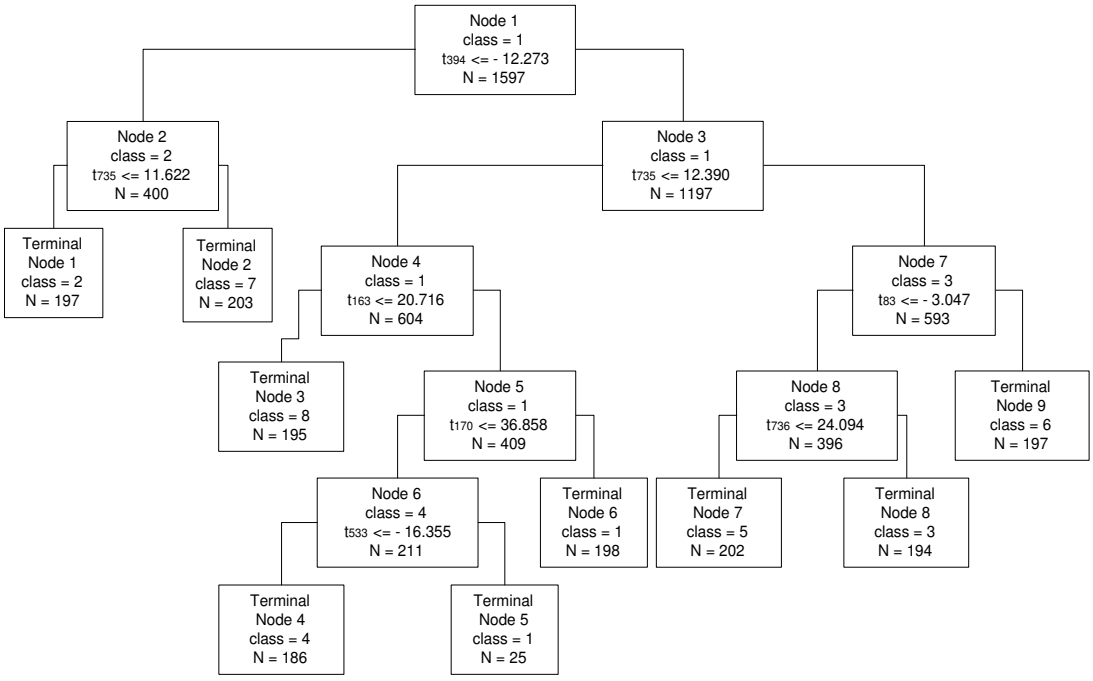

Figure 15: CART Tree in the Time Domain

Table 6: Misclassification error(%)

	Training data		Testing data	
Class	wavelet	time	wavelet	time
original	0.00	0.00	2.06	3.09
1	5.10	4.08	8.42	8.91
2	0.00	0.00	0.51	0.00
3	0.00	0.00	0.00	0.00
4	5.43	3.26	6.25	12.02
5	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.51
7	0.00	0.00	0.49	0.00
total error	1.25	0.87	2.25	3.13

Figure 15 shows the CART tree constructed using $N = 1,024$ points in the time domain. For this larger size data, it took 10 times more time to construct the decision tree than working with the reduced-size data set (5 versus 55 seconds in this small experiment). It took one second to obtain the reduced-size data set by applying the DWT and the RRE data reduction methods. In node 1, the first split is $t_{394} \leq -12.283$ where t_{394} is the value of the signal at time 394. In node 2, if $t_{735} \leq 11.622$, then the signal is classified into class 2; otherwise, the signal is classified as class 7.

The misclassification rates in the wavelet and time domains and in the training and testing samples are shown in Table 6. The CART tree in the time domain was almost perfect with respect to the training data, but it adapted too much to the features specific to the training data, and lost its generalization power. Thus, it did not work well when applied to the test data sets. CART with the reduced-size data is as comparable as the one obtained with the original larger size data in the training samples, but performs better in the testing samples. Overall, the total misclassification rate of CART for testing data set is 3.13% in the case with 1,024 data and 2.25% in the reduced-size data with 83 features. The existence of noise in signals makes classification in time domain difficult. Our *RRE*-based

methods reduce the data and remove the noise simultaneously for effective and efficient signal classification.

Remark: Our procedures were compared with the principal coordinates approach based on the function data-analytic method proposed in Hall and Poskitt (2001). Their method approximates the signal using the first M Karhunen-Loève basis functions with M decided from the cross-validation for minimizing the error in a specific decision method, e.g., CART classification in our application here. Although our data reduction methods are not designed for any specific decision method, for a comparison purpose, we found that CART's total misclassification rates from all eight data signal classes for their and our methods (e.g., RRE_s with $M = 83$ out of $1,024$ data points per curve) are 2.82% and 2.25%, respectively. Similar observations were obtained from normal distribution based quadratic discriminant analysis (QDA) advocated in Hall and Poskitt (2001), which has much higher total misclassification rates (about 25% in both methods). Because their method will require more computing effort, difficult to interpret the selected coordinates (in the sense of the reduced-size data), and might not be appropriate when the number of replicates is limited (smaller than L) and the data signal is noisy, our procedures are more useful in data reduction for various types of decisions.

2.7 Selection of Wavelet Positions Based on the Feature Selection Tool

In section 2.4, we developed data reduction procedures for generic purposes. Also, the best-basis algorithm selects wavelet coefficients suitable for signal compression (Coifman and Wickerhauser, 1992). For the classification problems, however, we can select those

wavelet coefficients that give large discrimination among the classes to reduce the rate of misclassification errors. In this section, we propose a method for selecting wavelet positions for signal classifications. This method, based on the feature selection tool, is useful for signal classification when the training data set is given.

In comparison with the enormous amount of attention devoted to signal analysis, compression and denoising, the wavelet transform has received relatively little attention as a basis for pattern recognition. The main feature of wavelets is that they are able to provide localized frequency information about a function or signal. Such information is particularly beneficial for signal classification. We can select basis functions that are well-localized in the time-frequency plane and that most discriminate between given classes.

Assume that a few classes of faults are considered. We limit our study to a few alternatives from known representative faulty processes. Several dissimilarity measures can be used to select wavelet positions, and evaluate the effectiveness of class discrimination (Fukunaga, 1990). Our procedure is based on a divergence measure because this measure is additive for independent variables. Since the wavelet coefficients are independent of each other in our model, the selection of wavelet positions can be done in a simple manner. Even though the noises are correlated in the time domain, the wavelet coefficients can be uncorrelated because of a decorrelating property.

Divergence is a measure of "distance" or dissimilarity between two classes. It can be used to determine feature ranking and to evaluate the effectiveness of class discrimination. Let the probability of occurrence of pattern $\boldsymbol{d}$, given that it belongs to class ω_i, be $p_i(\boldsymbol{d}) = p(\boldsymbol{d}/\omega_i)$, and the probability of occurrence of pattern $\boldsymbol{d}$, given that it belongs to class ω_j, be $p_j(\boldsymbol{d}) = p(\boldsymbol{d}/\omega_j)$. The divergence is defined as the total average information for

discriminating class ω_i from class ω_j, and given by

$$J_{ij} = \int [p_i(\boldsymbol{x}) - p_j(\boldsymbol{x}] \ln \frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})} d\boldsymbol{x}, \tag{16}$$

where $p_i(\boldsymbol{x})$ is the probability density function of class i.

Suppose that we have two signal classes characterized by two n-dimensional multivariate normal distributions: $N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i)$ and $N(\boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j)$, in which $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are the mean vectors, and Σ_i and $\boldsymbol{\Sigma}_j$ are $N \times N$ covariance matrices. The population densities are

$$p_k(\boldsymbol{d}) = \frac{1}{(2\pi)^{N/2}} |\boldsymbol{\Sigma}_k|^{1/2} \exp[-\frac{1}{2}(\boldsymbol{d} - \boldsymbol{\theta}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{d} - \boldsymbol{\theta}_k)], \ k = i \ or \ j. \tag{17}$$

The logarithm of the likelihood ratio in (16) is

$$\ln \frac{p_i(\boldsymbol{d})}{p_j(\boldsymbol{d})} = \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} - \frac{1}{2} \mathrm{tr}[\boldsymbol{\Sigma}_i^{-1} (\boldsymbol{d} - \boldsymbol{\theta}_i)(\boldsymbol{d} - \boldsymbol{\theta}_i)'] + \frac{1}{2} \mathrm{tr}[\boldsymbol{\Sigma}_j^{-1} (\boldsymbol{d} - \boldsymbol{\theta}_j)(\boldsymbol{d} - \boldsymbol{\theta}_j)'].$$

Hence, the divergence for these two classes is

$$J_{ij}(\boldsymbol{x}) = \frac{1}{2} \mathrm{tr}[(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1})] + \frac{1}{2} \mathrm{tr}[(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'].$$

In case of $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$, the divergence is

$$J_{ij}(\boldsymbol{d}) = \frac{1}{2} \mathrm{tr}[\boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'] = \frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)], \tag{18}$$

which equals a Mahalanobis generalized distance. For a univariate normal distribution, $n = 1$,

$$J_{ij}(\boldsymbol{d}) = \frac{(\theta_i - \theta_j)^2}{\sigma^2}, \tag{19}$$

in which θ_i and θ_j are the means, and σ^2 is the variance.

The divergence has the following useful properties.

1. $J_{ij} \geq 0$ (equality holds only when $i = j$).

2. $J_{ij} = J_{ji}$ (symmetric).

3. J_{ij} is additive for independent variables, i.e., $J_{ij}(d_1, d_2, \ldots, d_m) = \sum_{k=1}^{m} J_{ij}(d_k)$.

4. Adding a new measurement never decrease the divergence, i.e., $J_{ij}(d_1, d_2, \ldots, d_m) \leq J_{ij}(d_1, d_2, \ldots, d_m, d_{m+1})$.

The additive property of divergence is very useful when we use the wavelet transforms as a feature selection tool for signal classification. Because of the decorrelating property of the wavelets, the divergence based on p wavelet coefficients is equal to the sum of the p divergences based on each wavelet coefficient separately. This property may be used to determine the relative importance of each of the various features to be selected. The wavelet positions that will lead to a large divergence are the more important ones because they carry more discriminatory information. Thus, we may rank the importance of each wavelet coefficient according to its associated divergence. Any wavelet position that makes ONLY??? a small contribution to the total divergence may be discarded. The divergence concept provides us with a convenient way to order and select wavelet positions.

For a univariate normal distribution, say d_k, the divergence between class i and class j is obtained by

$$J_{ij}(d_k) = \frac{(\theta_{k,i} - \theta_{k,j})^2}{\sigma^2}$$

in which $\theta_{k,i}$ and $\theta_{k,j}$ are the means of class i and class j for variable d_k, respectively, and σ^2 is its variance. When there are L multiple classes in the training data set, we can define $J(d_k)$ as the sum of $\binom{L}{2}$ pairwise combinations of $J_{ij}(d_k)$.

When p wavelet coefficients are selected, the effectiveness measure may be determined by $\tau_{ij}(p)$. With an additional wavelet coefficient taken, the effectiveness measure is given by $\tau_{ij}(p+1)$. Then the incremental effectiveness that results from the addition of a wavelet coefficient is

$$\tau_{ij}(p+1) - \tau_{ij}(p).$$

Let the additional wavelet coefficient be d^*_{p+1}, which has mean θ^*_i or θ^*_j and variance σ^2; and let $\boldsymbol{z}$ be the vector covariance between d^*_{p+1} and the elements of $\boldsymbol{d}$. Then the new mean vectors and new covariance matrix are $\boldsymbol{\theta}^\nu_k = (\boldsymbol{\theta}_k; \theta^*_k)'$, $(k = i \text{ or } j)$ and

$$\boldsymbol{\Sigma}_{p+1} = \begin{pmatrix} \boldsymbol{\Sigma}_p & \boldsymbol{z} \\ \boldsymbol{z}' & \sigma^2 \end{pmatrix}.$$

The inverse of the new covariance matrix is

$$\boldsymbol{\Sigma}^{-1}_{p+1} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_p + \boldsymbol{\gamma}\delta^{-1}\boldsymbol{\gamma}' & -\boldsymbol{\gamma}\delta^{-1} \\ -\delta^{-1}\boldsymbol{\gamma}' & \delta^{-1} \end{pmatrix},$$

where $\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1}_p \boldsymbol{z}$ and $\delta = \sigma^2 - \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}_p \boldsymbol{z}$.

The effectiveness measure

$$\begin{aligned} \tau_{ij}(p+1) &= (\boldsymbol{\theta}^\nu_i - \boldsymbol{\theta}^\nu_j)'\boldsymbol{\Sigma}^{-1}_{p+1}(\boldsymbol{\theta}^\nu_i - \boldsymbol{\theta}^\nu_j) \\ &= [(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'(\theta^*_i - \theta^*_j)] \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_p + \boldsymbol{\gamma}\delta^{-1}\boldsymbol{\gamma}' & -\boldsymbol{\gamma}\delta^{-1} \\ -\delta^{-1}\boldsymbol{\gamma}' & \delta^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_i - \boldsymbol{\theta}_j \\ \theta^*_i - \theta^*_j \end{pmatrix} \\ &= \tau_{ij}(p) + \frac{1}{\delta}[(\theta^*_i - \theta^*_j) - (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'\boldsymbol{\gamma}]^2 \end{aligned}$$

Then, the incremental effectiveness due to the addition of a wavelet coefficient is given

by

$$\Delta_{ij} = \tau_{ij}(p+1) - \tau_{ij}(p) = \frac{[(\theta_i^* - \theta_j^*) - (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{z}]^2}{\sigma^2 - \boldsymbol{z}'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{z}} \tag{20}$$

When the additional wavelet coefficient d_{p+1}^* is uncorrelated with the other selected wavelet coefficients $d_1^*, \ldots, d_m^*$, we have this simple form:

$$\Delta_{ij} = \frac{(\theta_i^* - \theta_j^*)^2}{\sigma^2}. \tag{21}$$

Because the wavelet has a de-correlating property, we can easily select the good features that yield good discrimination among classes. And this approach will decrease the misclassification error rate. We can add the wavelet coefficients until the incremental effectiveness is less than a specified threshold value; the optimal threshold value can be determined based on a cross validation approach in which the criterion is the misclassification error rate.

Using the selected wavelet coefficients, linear discriminant analysis (LDA), quadratic discriminant analysis, artificial neural network analysis, and CART can be applied to classify the signals If the tree-based classification is combined with the wavelets capture of local information in the time-frequency plane, the interpretation of the result becomes clearly explicit and easy.

CHAPTER III

SPC PROCEDURES FOR COMPLICATED FUNCTIONAL DATA

3.1 Introduction

Recent years have witnessed rapid growth in data collection capabilities in engineering processes. Thus, utilizing functional or spatial data in quality improvement activities has become more feasible and widespread. For example, using spatial data from a semiconductor manufacturing process to detect process faults, Gardner *et al.* (1997) examined changes in signature patterns. Utilizing linear functional data, Lawless *et al.* (1999) monitored automotive manufacturing quality, and Kang and Albin (2000) monitored a mass flow controller in a semiconductor manufacturing process. Nair, Taam, and Ye (2002) analyzed functional data in robust design studies. Ganesa, Das, Sikder, and Kumar (2002) modeled acoustic emission signals to improve nano-machining process quality. Jin and Shi (1999) used tonnage signals to detect faults in a sheet-metal stamping process. Also, Lada, Lu, and Wilson (2002) analyzed quadruple mass spectrometry (QMS) samples of a rapid thermal chemical vapor deposition (RTCVD) process to detect significant deviations from the nominal process.

Researchers have proposed several statistical process control procedures that monitor linear functional data. For example, Kang and Albin (2000) proposed a multivariate T^2

method and a residual-based approach in building control charts for Phase I and Phase II analyses on linear functional data. This work was followed by Kim, Mahmoud, and Woodall (2003) and Mohmoud, and Woodall (2003), who improved procedures and carefully evaluated control-chart properties. Woodall, Spitzner, Montgomery, and Gupta (2003) provided a comprehensive review of this growing field. (Researchers use the average run length (ARL) to compare the performance of competing control chart methods. This ARL is a function of the power in a test statistic used to build the SPC charts.) See Woodall (2000) for details of the distinction between Phase-I and -II studies and basic SPC terms.

Unlike the linear functional data studied in current SPC research, this article focuses on "complicated" functional data observed in many real-life applications. Figure 1(a) gives an example of complicated functional data in Nortel's antenna manufacturing system. Because of the increasing popularity of wireless communications, the demand for antenna equipment to send and receive signals is growing rapidly. The technologically sophisticated antennae developed for this market require a high degree of quality during their production process. The testing equipment at Nortel receives antenna signals at different degrees of azimuth and elevation. For the purposes of detecting process fault(s) quickly, engineers developed a heuristic monitoring procedure based on the central azimuth curve as shown in Figure 1(b). For example, the three main lobes in the center provide the most important information for typical usage. Certain specification limits are set on the peaks and amongst the differences between the peaks and valleys.

Many other examples of complicated functional data exist in a wide range of applications. See Jin and Shi (1999), Bakshi (1999) and Ganesan *et al.* (2002) for examples.

Typically these complicated functional data have nonlinear patterns with many local sharp-changes providing important process information. Moreover, possible dependence between successive data points and potentially large size data sets (e.g., $n = 256$ in Figure 1(b)) make multivariate data analysis difficult.

In this article, we focus on Phase-II analysis with the goal of detecting process problems quickly using the new data and the baseline model established in Phase-I studies. In the Phase-II research of linear profile data, model parameters such as the intercept, slope, and error variance are monitored (e.g., Kim, Mahmoud, and Woodall (2003)) for detection of a possible change in their sizes. When dealing with complicated functional data, one approach is to extend their ideas by using a higher-order polynomial or nonlinear regression to model the data and, then, monitor key model parameters representing data trends. There are several challenges in this approach, especially in dealing with the data illustrated in Figure 1. First, the regression models and even Fourier transforms do not perform well in modeling sharp changes. Evidences are given in Jin and Shi (2001), Ganesan *et al.* (2002), where wavelet transforms were advocated by these authors. See Section 2 for a brief review of wavelet transforms.

A more important challenge is that too many parameters are monitored when we fit a model to complicated functional data. It is well known (e.g., Fan, 1996) that the power of detecting process faults will drop significantly when the size of parameter vector becomes large. This implies that the ARL will become very large. Functional principal component analysis (FPCA; Ramsay and Silverman, 1997) and related procedures (see Hall, Poskitt, and Presnell (2001) for an example) are useful in modeling nonlinear profile data. It can be used as a dimension-reduction tool for handling the "power-drop" problem. For example,

Jones and Rice (1992) used principal component analysis to identify and illustrate important modes of variation among several curves. However, there are some difficulties in applying the FPCA approach to solve the SPC problems. For instance, FPCA lacks interpretation ability where the relationship between changes of the selected principle components and functional data is unclear. More importantly, its ability to model sharp changes and detect local shifts is doubtful. Thus, wavelet transforms will be our main modeling procedure in this article.

To reduce the size of wavelet model-parameters, Jin and Shi (1999) utilized engineering knowledge to select a few wavelet coefficients for monitoring. Jin and Shi (2001) used a data denoising technique (see Donoho and Johnstone (1994) for details) to select several wavelet coefficients and apply the multivariate analysis approach based on the Hotelling T^2 statistic to detect process faults. Jeong, Chen, and Lu (2003) presented a thresholded scalogram approach to monitor process changes. All methods outlined above first apply a "feature-selection" tool in the wavelet domain to reduce the dimension of functional data. Then, they construct an appropriate test statistic based on the selected wavelet-features. There are many concerns with this approach. First, depending on the level of noise in the data, the number of selected features based on the popular thresholding procedures (e.g., Donoho and Johnstone, 1994 and 1995) may still be large. Although Jeong, Lu, Huo, Vidakovic, and Chen (2002) developed a procedure to limit the size of these features, the objective in the feature-selection process is not to minimize the ARL. Moreover, the next few paragraphs show that because the selected wavelet-features for monitoring are fixed (based on the in-control baseline data and some known types of process faults), these approaches are not effective in detecting faults that lead to changes in the unselected wavelet-features.

The current SPC procedures for the linear profile data in the literature are focused on detecing "global shifts" patterns that change the entire profile. For example, in detecting possible process changes in mean linear regression function, Mahmoud *et al.* (2003) proposed Phase-II EWMA procedures to monitor the intercept and slope parameters separately. The SPC limits involve known model parameters established in Phase-I analysis. Their ARL performance evaluations focused on shifts of these parameters from the known values in the in-control situation to increments of them in terms of some units of the standard deviation. Any change of the two regression parameters will lead to a change in the entire linear profile data.

In contrast to typical "global shift" studies, our current research focuses on "local shifts." As an illustrating example, assume that a small percentage (e.g., 5%) of data in the middle of a linear profile-data all increased up by a certain unit of the standard deviation. Depending on the amount of shift, the fitted linear regression line from new data may not be very different due to the "averaging" effect in the estimates of regression parameters. Thus, the EWMA charts might not be able to detect these local shifts. This means that using only the two regression parameters "selected" from analyzing baseline process data may not be able to detect local shifts.

The local shifting problem is of major significance in our study. For example, some changes in valleys or peaks of Figure 1 can present a major quality problem in the antenna manufacturing process. If the SPC charts are built on parameters selected using baseline data and the new process data have a shift in a local segment such that the selected parameters cannot characterize this shift, the power of detecting this type of process change will be very low and the parameter size m could be large. Furthermore, changes at several

process-runs can be very different. This leads to distinct patterns in local changes. Thus, the selected parameters need to be updated based on the difference between the new data and the baseline process information. However, in this article, our procedure will monitor the sum of selected parameters similar to the Hotelling T^2 statistic. Although the parameters to be included in the sum are different for every new data set, the functional form of the monitoring statistic remains the same. This will somewhat ease the use of an "unconventional" procedure in handling many possible local-shifting problems.

3.2 Problem Formulations

3.2.1 Wavelet Approaches

Wavelet transforms can model irregular data patterns such as sharp changes in Figure 1 better than the Fourier transforms and standard statistical procedures (e.g., parametric and nonparametric regressions) and provide a multi-resolution approximation to the data (Mallat, 1989). Wavelet transforms have been demonstrated to be effective in audio and image processing applications (e.g., Rao and Bopardikar, 1998; Chapter 5) and many data-denoising studies (e.g., Donoho and Johnstone, 1994). Rying, Bilbro, and Lu (2002) used it to extend the ability of the artificial neural network (ANN) in learning complicated data patterns with local focus. See Lada *et al.* (2002), Jeong *et al.* (2002), and Jeong *et al.* (2003) for applications in detecting manufacturing anomalies.

A wavelet is a square integrable function with a zero average and unity norm. Wavelets can be translated (u) and dilated (s) to create a family of time-frequency atoms, $\phi_{s,u} = \phi[(t-u)/s]$. An example of the $\phi(t)$ function is the "sombrero" wavelet (see page 77 of Mallat, 1998), which looks like a Mexican sombrero. If $f(t)$ is also square integrable, then

$f(t)$ can be expressed (Daubechies, 1992) as the following equation:

$$f(t) = \sum_{k \in Z} c_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{\infty}\sum_{k \in Z} d_{j,k}\psi_{j,k}(t), \quad Z = 0, \pm 1, \pm 2, \tag{22}$$

where $\phi_{L,k}(t)$ ($\psi_{j,k}(t)$) are the father (mother) wavelets representing the low-frequency and smooth (high-frequency and detail) parts of a signal. The wavelet coefficients $c_{L,k}$ and $d_{j,k}$ are defined as inner products of $f(t)$ and the corresponding wavelet functions, $\phi_{L,k}(t)$ and $\psi_{j,k}(t)$, respectively. In practice, the following finite version of the wavelet series approximation is used:

$$\tilde{f}(t) = \sum_{k=0}^{2^L-1} c_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{J}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t)\,, \tag{23}$$

where $2^J = n$ and L corresponds to the lowest decomposition level.

Follow the traditional model used in statistical studies of wavelets. Suppose that when a process is statistically controlled, the functional data collected over time can be represented as follows:

$$Y(t_i) = f_0(t_i) + \epsilon(t_i),\ i = 1, 2, \ldots, n, \tag{24}$$

where $f_0(t_i)$ is the known target-signal established in Phase-I studies, and $\epsilon(t_i)$'s are independent and identically distributed (i.i.d.) normal random variables with mean zero and variance σ^2. Let $\boldsymbol{Y}$, $\boldsymbol{f}_0$ and $\boldsymbol{\epsilon}$ be the collections of $Y(t_i)$'s, $f(t_i)$'s and $\epsilon(t_i)$'s at n equally spaced time points.

The discrete wavelet transform (DWT) of $\boldsymbol{y}$ is defined as

$$\boldsymbol{d} = \boldsymbol{W}\boldsymbol{y}, \tag{25}$$

where $\boldsymbol{W} = [h_{ij}]$, for $i, j = 1, 2, \ldots, n$ is the orthonormal $n \times n$ wavelet-transform matrix. The matrix $\boldsymbol{W}$ is different according to the wavelet type, the decomposition level, and the number of sample points n. The elements h_{ij}'s have a special structure, corresponding to a sequence of linear filtering operations. In practice, the pyramid algorithm is used to compute the wavelet and inverse wavelet transforms in $O(n)$ operations (Mallat, 1989). The DWT transforms n data points into n wavelet coefficients and is computationally (with their $O(n)$ calculation complexity) superior to any other signal processing or statistical modeling procedures. For example, Fourier transforms possess complexity $O(n \log n)$ and the PCA requires solving an eigenvalue system which is an expensive $O(n^3)$ operation. If $\boldsymbol{W}$ is orthonormal, the original data $\boldsymbol{Y}$ can be reconstructed by the inverse DWT as $\boldsymbol{y} = \boldsymbol{W}^{-1}\boldsymbol{d}$. The tremendous practical success of wavelets is based on their ability to parsimoniously represent the model of data by only a few important wavelet coefficients.

Apply the DWT to the data $\boldsymbol{y}$ of random variables $\boldsymbol{Y}$, and obtain the following wavelet coefficients:

$$\boldsymbol{d} = \boldsymbol{\theta}_0 + \boldsymbol{\eta}, \tag{26}$$

where $\boldsymbol{\theta}_0 = \boldsymbol{W}\boldsymbol{f}_0$, and $\boldsymbol{\eta} = \boldsymbol{W}\boldsymbol{\epsilon}$ is $N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ distributed (Vidakovic, 1999, page 169), where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Let us use the following theorem to better understand the relationship between the mean $\boldsymbol{f}$ and its DWT $\boldsymbol{\theta}$ in the case where a local segment of data is shifted. By analyzing the changed wavelet coefficients the result (ii) can be used to identify the locations of changes in the original time domain. This "mapping" property can facilitate the search of process faults and their causes. See Appendix for its proof.

Theorem 4.

(i) When there is a process mean shift in ρ_i units of σ for the time t_i's in the interval $A = (t_s, t_e)$ with $t_1 \leq t_s < t_i < t_e \leq t_n$, i.e.,

$$f_{new}(t_i) = \begin{cases} f_0(t_i) + \rho_i\sigma, \ t_i \in A, \\ f_0(t_i), \ elsewhere, \end{cases}$$

the true wavelet coefficients have a corresponding shift given as follows:

$$\theta_{i,new} = \theta_{i,0} + \delta_i\sigma, \ i = 1, 2, \ldots, n,$$

where $\delta_i = \sum_{j\in A} \rho_j h_{ij}$.

(ii) When there is a parameter shift in γ_i units of σ in the wavelet coefficients from the "area" B, i.e.,

$$\theta_{i,new} = \begin{cases} \theta_{i,0} + \gamma_i\sigma, \ i \in B, \\ \theta_{i,0}, \ elsewhere, \end{cases}$$

the mean function in the time domain will have a corresponding shift as follows:

$$f_{new}(t_i) = f_0(t_i) + \tau_i\sigma, \ i = 1, 2, \ldots, n,$$

where $\tau_i = \sum_{k\in B} \gamma_k h_{ki}$.

Remark 3.1 For a vertical shift of the entire data curve, the process mean is shifted vertically, in Theorem 1 $t_s = t_1$, and $t_e = t_n$ with $\rho_i = \rho$, $i = 1, \ldots, n$. Then, $\delta_i = \rho \sum_{j=1}^{n} h_{ij}$. In the case of Haar wavelet with the decomposition level L in Equation (1), $\sum_{j=1}^{n} h_{ij} = 2^{(J-L+1)/2}$, for $i = 1, 2, \ldots, 2^L$; zero, for other i's. The first $i = 1, 2, \ldots, 2^L$

wavelet-coefficients belong to the coarser level (father wavelets) and the rest coefficients are for the finer levels. From this property of haar wavelet, the true wavelet coefficients have the following shift in the case of the vertical shift: $\theta_{i,new} = \theta_{i,0} + 2^{(J-L+1)/2}\rho\sigma$, for the coarse level $i = 1, 2, \ldots, 2^L$; and $\theta_{i,new} = \theta_{i,0}$, for the other i's in the finer levels. That is, the wavelet coefficients in the finer levels *do not change*.

Remark 3.2 For local-segment shifts, the baseline signal has been changed for some local area. In this case, the wavelet coefficients which have common support from t_s to t_e will be affected. For example, if the baseline signal from $t_s = 1$ to $t_e = 3$ (with $n = 128$ and $L = 3$) has been changed with $\rho_i\sigma$ level, then DWT-coefficients $c_{3,1}$ in the coarser level and $d_{3,1}, d_{4,1}, \ldots, d_{J,1}$ and $d_{J,2}$ in the finer levels will be changed.

3.2.2 Problem Formulations

Formulating process-monitoring procedures for complicated functional data starts with understanding the following hypothesis-testing problem: for a new set of data $\boldsymbol{Y}_{new}$ from the $N_n(\boldsymbol{f}_{new}, \sigma^2 I_n)$ distribution, test

$$H_0 : \boldsymbol{f}_{new} = \boldsymbol{f}_0 \text{ versus } H_1 : \boldsymbol{f}_{new} \neq \boldsymbol{f}_0. \tag{27}$$

In the DWT-based wavelet domain, the above hypotheses become

$$H_0 : \boldsymbol{\theta}_{new} = \boldsymbol{\theta}_0 \text{ versus } H_1 : \boldsymbol{\theta}_{new} \neq \boldsymbol{\theta}_0. \tag{28}$$

It is known that the uniformly most powerful invariance (UMPI) test for (28) is based on a Hotelling T^2 statistic. For uncorrelated noises with a known variance parameter, the

Hotelling T^2 is equivalent to the χ^2-test given by the following equation:

$$\chi_0^2 = \sum_{j=1}^{n} \frac{(d_{j,new} - \theta_{j,0})^2}{\sigma^2}. \tag{29}$$

Thus, the following upper control limit (Montgomery, 2001) based on the chi-square distribution can be used to monitor potential process changes

$$UCL_1 = \chi_{\alpha,n}^2, \tag{30}$$

where $\chi_{\alpha,n}^2$ is the upper α percentage point of the chi-square distribution with n degrees of freedom.

When the dimension of the data n is large, the power of the χ^2-test can be unsatisfactorily low (Fan, 1996), which will lead to a very large average run length (ARL_1) when the process is out-of-control. For example, at $\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_1$ in H_1, the χ^2-test has the following approximated power:

$$1 - \Phi(\frac{z_{1-\alpha} - ||\boldsymbol{\theta}_1||^2/\sqrt{2n}\sigma^2}{\sqrt{1 + 2||\boldsymbol{\theta}_1||^2/(n\sigma^2)}}) \approx 1 - \Phi(z_{1-\alpha} - ||\boldsymbol{\theta}_1||^2/\sqrt{2n}\sigma^2).$$

This power tends to α even though $||\boldsymbol{\theta}_1||^2$ goes to infinity (with $||\boldsymbol{\theta}_1||^2 = o(\sqrt{n})$).

To overcome this difficulty, several authors proposed testing a subset of the coefficients. For example, Kasashima, Mori, Ruiz, and Taniguchi (1995), Mori, Kasashima, Yoshioka, and Ueno (1996) and many others used their "engineering knowledge" to decide which few wavelet coefficients to monitor for detecting a few known faults in manufacturing processes. Jin and Shi (2001) first applied the *VisuShrink* data denoising procedure (Donoho and Johnstone, 1994) to screen out smaller wavelet coefficients, which are viewed as unimportant coefficients for process monitoring. Then, they used the Hotelling T^2 with screened

important coefficients to develop a SPC procedure:

$$T_0^2 = \sum_{j \in S} \frac{(d_{j,new} - \theta_{j,0})^2}{\sigma^2}, \tag{31}$$

where S is the set of pre-selected wavelet coefficients by $VisuShrink$. Then, the control limit becomes

$$UCL_2 = \chi^2_{\alpha,p}, \tag{32}$$

where p is the number of pre-selected wavelet coefficients in S. However, it is possible that the subset of the coefficients monitored does not show any significant difference from the target in H_0, but other coefficients not monitored show significant difference. This means that we also need to monitor other coefficients to make sure that they are unchanged. To overcome this problem, we propose in Section 4 a procedure that considers only wavelet coefficients that deviate significantly from target values of wavelet coefficients adaptively depending on the change of data in process runs.

3.2.3 Other Options of Problem Formulations

3.2.3.1 Simple alternative hypothesis

In some applications, we are interested in only a specific type of fault (e.g., vertical shift or some meaningful local changes: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$). In this case, the problem can be formulated as follows:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1. \tag{33}$$

Given a specific alternative $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, the Neyman-Pearson fundamental theorem states that the most powerful test for the problem (33) is to reject H_0 when $(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^T(\boldsymbol{d}_{new} - \boldsymbol{\theta}_0) >$

$||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0||\sigma\Phi^{-1}(1-\alpha)$, where α is the significance level, Φ is the standard normal distribution function, and $||\cdot||$ is the L_2 norm. The power of this optimal test is $1-\Phi(z_{1-\alpha}-||\boldsymbol{\theta}_1-\boldsymbol{\theta}_0||/\sigma)$. In this case, we can get the best power even for the sparse cases without reducing the dimension. Thus, we do not need to apply thresholding techniques to overcome the problems caused by the high dimension.

Another application is when it is necessary to detect a random alteration. In this case, the problem can be formulated as follows:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\delta}, \tag{34}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p, 0, \ldots, 0)$. For this kind of problem, the likelihood ratio test was developed by Srivastara and Worsley (1986). The detailed procedure is skipped. Note that the difference between (33) and (34) is the covariance structure. In (33), we assume i.i.d. random noise with equal variance σ^2 while in (34) we assume the general covariance matrix Σ which is unknown.

3.2.3.2 SPC for restricted alternatives

Before moving to the proposed methods, let us discuss a few other alternatives from the literature in formulating the hypothesis-testing problems for establishing the SPC limits. Several researchers proposed solutions for testing hypotheses with the following restricted alternative:

$$H_0 : \boldsymbol{\theta}_{new} = \boldsymbol{\theta}_0 \text{ versus } H_1 : \boldsymbol{\theta}_{new} \geq \boldsymbol{\theta}_0.$$

By restricting the alternative to a subset of the general alternative in (28), one hopes that the power of the test could be improved, and thus ARL_1 would be reduced. Kudo

(1963) and Perlman (1969) developed likelihood ratio tests for the restricted hypotheses. However, due to the complicated distribution of the test statistic under H_0, these tests cannot be easily implemented in SPC applications. Tang (1994), Silvapulle (1995), and Wang and McDermott (1998) proposed the uniformly most powerful (UMP) test for this type of restricted hypotheses. However, the computation of its power is very difficult and time consuming for SPC implementation. Because of its complicated power function, there is no literature about the relationship between the dimension n and its power.

3.2.3.3 SPC with k known fault classes

With the same motivation as above, one could restrict the alternatives to a few fault classes. For example, consider that $\boldsymbol{\theta}_{b,a}$ (for $b = 1, 2, ..., k_b$) are (size-n) coefficients from k_b fault classes. One way to combine these multiple alternatives, which consist of coefficients larger or smaller than the target values, is to use a weighting function. For instance, $\boldsymbol{\theta}_a = \sum_{b=1}^{k_b} p_b \, \boldsymbol{\theta}_{b,a}$ with $\sum_{b=1}^{k_b} p_b = 1$. Thus, the hypotheses in testing become $H_0 : \boldsymbol{\theta}_{new} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta}_{new} = \boldsymbol{\theta}_a$, which is a simple hypothesis testing problem. In this case, the Neyman-Pearson theorem states that the most powerful test is to reject H_0 when $(\boldsymbol{\theta}_a - \boldsymbol{\theta}_0)^T(\boldsymbol{d}_{new} - \boldsymbol{\theta}_0) > \sigma||\boldsymbol{\theta}_a - \boldsymbol{\theta}_0||\Phi^{-1}(1-\alpha)$, where $||\cdot||$ is the L_2 norm. The power of this optimal test is $1 - \Phi(z_{1-\alpha} - ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0||/\sigma)$. Note that in this function the power will not be decreasing due to the increasing dimension n (i.e., we can test all coefficients). However, this approach has several concerns such as what fault classes should be included, how to assign their weights, the power might not be satisfactory (i.e., the ARL_1 could be reasonably small), and what if there are new faults not considered. Thus, this approach will not be discussed here, but instead, left to future work.

3.2.3.4 Step-down procedure

The Hotelling's T^2 procedure, which is the uniformly most powerful invariant (UMPI) test for (28) treats all variables symmetrically and may not be appropriate if the variables are of unequal importance. The wavelets has multi-resolution property and the importance of each resolution (or level) can be different. For example, the wavelet coefficients at the coarser level capture the global pattern of a signal. Thus, well-known denoising techniques such as $VisuShrink, RiskShrink$, and $SURE$ don't shrink the wavelet coefficients at the coarser level even though their values are small. That is, the information at the coarser level can be treated as more important than other levels from the aspect of fault detection.

A step-down procedure was proposed by Roy (1978) to consider the relative importance of each variable. However, this procedure is not appropriate for high-dimensional testing problems such as process monitoring of functional data. In high-dimensional testing problems, the type I error of each variable in the traditional step-down procedure should be almost zero to satisfy a global type I error α.

We present the SPC procedure, which combines the multi-resolution property of wavelets and the step-down procedure. In the wavelets, we have $J - L + 1$ number of resolutions when the dimension of the data is $n(= 2^{J+1})$ and the lowest decomposition level is L. Control charts can be established for each resolution level. Resolution levels (or scales) can be ordered according to their importance, and the overall type I error probability can be distributed suitably between the component tests. The null hypothesis H_0 is rejected as soon as a component test in the sequence shows significance.

We have $N(> J - L + 1)$ random samples and the test statistic is given as follows:

$$F_i = (n-i)(T_i^2 - T_{i-1}^2)/[(N-1) + T_{i-1}^2],\ i = 1, \ldots, J - L + 1$$

Under H_0 in (28), F_i's are independently distributed according to F distributions with degrees of freedom 1 and $N - i$.

We are dealing with single measurement. We have to develop procedures for the following cases: (i) known Σ, (ii) unknown Σ with N preliminary samples, and (iii) uncorrelated noise. The traditional step-down procedure was developed for single variables. However, we are dealing with multi-resolution and we have several variables at each multi-resolution. Consequently, we must develop a procedure to encompass several variables instead of a single one.

3.2.4 Adaptive Thresholding Hypothesis-testing Procedures and SPC Limits

Unlike all other procedures, which only monitor selected wavelet coefficients using baseline signals, our method selects wavelet coefficients adaptively using new data coming from recent process runs. This approach can prevent from having low detection probability for high-dimensional data and not monitoring unselected coefficients based on the baseline data.

Let us start with modifying the UMPI test given in (29). When a "hard-thresholding" (see Donoho and Johnstone (1994) for its detailed definition) is applied to the difference $|d_{j,new} - \theta_{j,0}|$ to retain only larger values of the difference (in the units of σ), the chi-square version of the UMPI test becomes

$$T_A^2 = \sum_{j=1}^{n} \frac{(d_{j,new} - \theta_{j,0})^2}{\sigma^2} I(|d_{j,new} - \theta_{j,0}| > \delta\sigma). \tag{35}$$

This test statistic is a modified version of Fan's hard-thresholding procedure for testing

significant difference between two curves (Fan, 1996). We took the difference between the new wavelet coefficients and target wavelet coefficients, and then standarized them using the variance. Here, we select wavelet coefficients by considering the information of process parameters. Only the wavelet coefficients which are deviated from target parameters are used to calculate the proposed test statistics and this can avoid low detection probability for high-dimensional data. Moreover, we select wavelet coefficients adaptively according to process changes, i.e., the selected wavelet coefficients can be different according to process faults and this approach will be very powerful in detecting new types of faults quickly.

The exact formula for the mean and variance of the proposed statistic T_A^2 are given by (see Jeong (2004) for details)

$$\mu_{n,H_0} = n[2\delta\phi(\upsilon) + 2(1 - \Phi(\delta))],$$

$$\sigma^2_{n,H_0} = 2\phi(\delta)n[\delta^3 + \delta^2(1 - 2\phi(\delta)) - 4\delta(1 - \Phi(\delta))],$$

where ϕ is the probability density function of a standard normal distribution. Fan (1996) derived the approximated formula for the mean and variance when *delta* is large. Although his formula is easier to implement, it is not appropriate for the following SPC applications, where smaller value of *delta* is needed for assuring the asymptotic normality (when n goes to infinity).

The exact distributions of statistics such as T_A^2 for monitoring complicated function data are intractable, and simulations of their finite-sample distribution are tedious and case- and parameter-specific. This article utilizes the normal distribution obtained from large-sample approximation theory (see Theorem 5) for building SPC limits. Thus, the upper control

limit based on the approximated distribution of T_A^2 is as follows:

$$UCL_3 = \mu_{n,H_0} + \sigma_{n,H_0}\Phi^{-1}(1-\alpha). \tag{36}$$

The quality of the asymptotic normality, however, depends on the dimension of data (n) and a threshold parameter (δ). Based on our experiments, for $n = 256$, the largest threshold parameter that makes asymptotic normality plausible under H_0 is 2.8. Thus, we restrict the range of δ from 0 (no thresholding) to 2.8 for $n = 256$. Under H_1, the performance of asymptotic normality of T_A^2 depends on the shift level of a process (or new mean of a process, $\boldsymbol{\theta}_1$), threshold parameter, and the dimension of data. The quality of the asymptotic normality becomes better under H_1 with a wider range of δ because the deviates of wavelet coefficients from the parameters given in H_0 are larger.

The performance of an adaptive thresholding test depends on the threshold value (δ). One approach is to find a thresholding parameter by maximizing the power of the proposed test based on the large sample distribution of the test statistic (35). Figure 16 shows an example of its power function plotted against *delta* using the case with a "local change" with shift level 1.2, where the range of δ is restricted to make the asymptotic normality plausible. The numerical value of optimal δ can be obtained from an algorithm based on the golden search and parabolic interpolation (see Forsythe, Malcolm, and Moler (1976) for details).

Theorem 5. Consider a shift level $\omega = (\omega_1, \ldots, \omega_n)$ where $\omega_j = |\theta_{j,1} - \theta_{j,0}|/\sigma, j = 1, \ldots, n$ and $\theta_{j,1}$'s are new process mean parameters. Let $X_j = [(d_{j,new} - \theta_{j,0})^2/\sigma^2]I(|\tau_j| > \delta)$. Assume $\mu_{n,H_1} = \mathrm{E}(\sum_{j=1}^n X_j|\omega_i; i = 1, \ldots, n) \geq 0$ and assume that $\sigma_{n,H_1}^2/n = Var(\sum_{j=1}^n X_j|\omega_i; i = 1, \ldots, n)/n \to \sigma^2$ as $n \to \infty$. Then, the asymptotic distribution of T_A^2 under H_1 ($: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$)

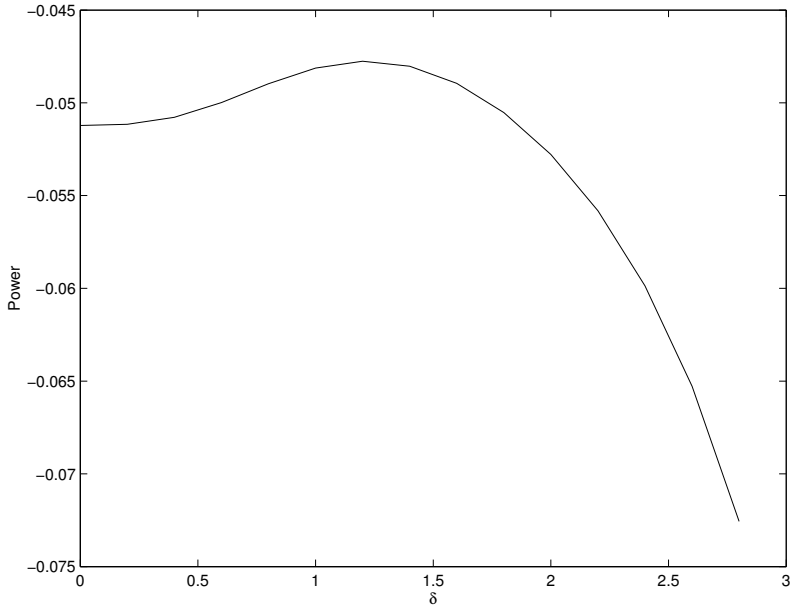

Figure 16: Power Function Under Local Shift ($n = 256$)

is,

$$\frac{T_A^2 - \mu_{n,H_1}}{\sqrt{n}\ \sigma} \xrightarrow{d} N(0,1) \quad as \quad n \to \infty.$$

Given a new process mean $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, the asymptotic power of T_A^2 is a function of $(\boldsymbol{\theta}_1)$ and threshold parameter (δ), and is calculated by

$$1 - \Phi\left(\frac{\mu_{n,H_0} - \mu_{n,H_1} + \sigma_{n,H_0} z_{1-\alpha}}{\sigma_{n,H_1}}\right). \tag{37}$$

See Jeong (2004) for the complicated formulas of the mean and variance of T_A^2 under H_1. Table 7 compares the powers of different procedures in case of local changes with shift level up to 1.0. As the shift level gets large, the powers of all procedures increase and the SPC performs better. Thus, the results of large size shifts are not presented here. For T_A^2, the power is given for the optimal threshold parameter. This result shows that we can get better performance (higher power or low ARL_1) by selecting the value of threshold parameter based on the power function than using other statistics.

Table 7: Comparison of Powers of different procedures

shift level	0.4	0.6	0.8	1.0
χ^2	0.0094	0.0194	0.0474	0.1227
T_0^2	0.0133	0.0384	0.1268	0.3690
T_A^2	0.0258	0.1581	0.3611	0.5145

Table 8: UCL Values vs Threshold Values

δ	0.0	0.5	1.0	1.5	2.0	2.5
μ_{H_0}	256.00	248.09	205.12	133.67	66.93	25.61
σ_{H_0}	22.62	22.94	23.97	23.55	20.11	14.64
UCL	308.63	301.47	260.89	188.48	113.71	59.68

Remark 3.3 The values of threshold parameter in an adaptive thresholding procedure can be changed at every monitoring points according to data. Fixing δ at every monitoring point may not be good at detecting various kinds of shifts. The optimal threshold can be different according to new mean $\boldsymbol{\theta}_1$, (i.e., according to the the shift levels and shift types), resulting in varying UCL values. A higher threshold value gives smaller UCL. Table 8 shows some examples of values of mean and standard deviation under H_0, and UCL values according to values of threshold in case of $n = 256$ and $\alpha = 0.01$. The values of optimal threshold could be different under different situations (various shift levels and shift types), therefore the values of UCL could be different under different situations.

In practice, the mean of changed process ($\boldsymbol{\theta}_1$), however, is unknown and we have to estimate it based on the observed data. Let $\omega_j = |\theta_{j,1} - \theta_{j,0}|/\sigma, j = 1, \ldots, n$ be the standarized shift level of the process when the process is out-of-control ($\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_1$). The naive estimate of this shift level based on observed data $\boldsymbol{d}_{new}$, is $\tau_j = (d_{j,new} - \theta_{j,0})/\sigma, j = 1, \ldots, n$ and τ_j's are i.i.d. Gaussian random noises under H_0 ($\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_0$). We can improve the quality of the estimate using James-Stein estimate (Stein, 1981) by reducing the impact of process noises. It is reasonable to assume that only a few τ_j's contain information about

the real process-shift while others are contaminated by random noises. The goal is to extract these significant coefficients and to ignore others. Such an extraction can be naturally performed by thresholding the τ_j's. This leads to the J-S estimate of true process shift is $\hat{\omega}_j = \tau_j I(|\tau_j| > \lambda_0)$, where λ_0 is a well-known global threshold. The commonly used data-denoising method, MinMax threshold, can be applied (Donoho and Johnstone, 1995).

Based on the estimate of new process mean, we can calculate the approximated power function and get the threshold value which maximizes the power function in a similar way. In this case, the power is the function of the estimate of a new process mean ($\hat{\boldsymbol{\theta}}_1$) and threshold parameter (δ). We call this procedure T^2_{B1}.

Another approach is that the thresholding parameter can be found from the following equation, modifying the idea suggested by Fan's procedure,

$$\delta = \sqrt{2\log(n\hat{a}_n)}, \qquad \hat{a}_n = \min(4(\max_{1\le i\le n}(d_{i,new} - \theta_{i,0})/\sigma)^{-4}, \log^{-2} n). \tag{38}$$

We call this procedure T^2_{B2}. Our preliminary experience (see Tables 9, 10, and 11) indicates that when process-noises are involved, the above procedure (38) could have poor ARL_1 for smaller process-shifts. One possible reason is that this thresholding is similar to the rules used for data-denoising purposes.

Note that wavelet coefficients to be monitored in both T^2_{B1} and T^2_{B2} will be changed according to the data set adaptively while they are pre-determined and fixed in T^2_0 in (31). When process changes occur, the wavelet coefficients ($d_{j,new}$'s) will deviate from the target values ($\theta_{j,0}$'s) and the number of wavelet coefficients to be included in T^2_{B1}(T^2_{B2}) will increase, resulting in a larger value of $T^2_{B1}(T^2_{B2})$ so that H_0 is rejected (out-of-control). The performance of these procedures will be compared in the next section.

3.2.5 Simulation Studies

This section presents simulation results that compare the ARL_1 from the SPC limits based on the following four test-statistics: UMPI χ_0^2, $VisuShrink$-based T_0^2, T_{B1}^2, and T_{B2}^2. Assume that wavelet coefficients of a baseline signal, $\boldsymbol{\theta}_0$, are known for the Phase II monitoring-charts. But, the parameter $theta_1$ for the possible changed process in H_1 is unknown and estimated from the data. For all the SPC charts the parameter that determines the control limit from the center line is set so that the in-control ARL (ARL_0) is equal to 200, a typical number used in SPC studies (Mahmoud and Woodall, 2003).

To validate that our procedures have the ability to handle sharp changes in data curves, in this simulation study, we use the antenna signal presented in Figure 1 (b) with $n = 256$ as a underlying mean curve. Random noises from normal $N(0, \sigma^2)$ with $\sigma^2 = 1$ are added to generate 1,000 replications for each study in Tables III, IV and V. Three types of shifts are considered: (1) three local segments at intervals [5, 7], [80, 85] and [240, 243] are shifted (for a total of 13 out of 256 data points (5%)), (2) shift of a center-segment [123, 133], which has 4.2 % (11 data points) of the whole data set, and (3) vertical shift, where the new curve is vertically shifted in the entire range from the original one by $\gamma\sigma$. Two wavelet families (Haar and Symmlet-8) are used for examining their effects and similarities. Because the results from these two wavelet families are similar, only results with Symmlet-8 wavelets are presented here. For all studies the lowest decomposition level (L) is set to 4.

3.2.5.1 Local Changes

Table 9 and Figure 17 give the ARL_1 values. Both the T_{B1}^2-chart and the T_{B2}^2-chart perform better than the UMPI χ_0^2-chart and the T_0^2-chart over the entire range of the shifts tested.

Table 9: Comparison of ARLs Under Local Shifts ($\gamma\sigma$)

	γ					
Chart	1.2	1.4	1.6	1.8	2.0	2.2
χ_0^2	23.04	13.80	7.20	4.92	2.69	2.00
T_0^2	64.07	47.99	33.11	24.07	13.17	8.89
T_{B1}^2	13.98	7.89	4.94	3.30	1.82	1.64
T_{B2}^2	17.31	10.28	6.01	3.73	2.21	1.78

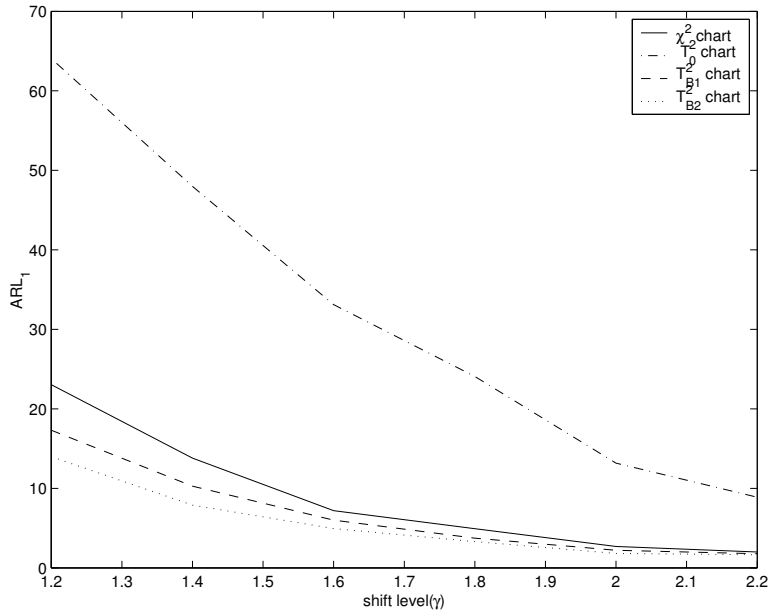

Figure 17: ARL Comparisons Under Local Shifts

Because the T_0^2-chart fixes the wavelet coefficients to be monitored, it is not sensitive to local changes. Both the T_{B2}^2 chart and the T_{B1}^2 chart are adaptive to process shifts and consider only those wavelet coefficients that undergo large changes. The T_{B1}^2-chart performs better than the T_{B2}^2-chart over the entire range of shifts considered. Compared to other procedures, the T_{B1}^2-chart effectively removes the noise in the estimation of the shift-information and works well given the lack of any prior information indicating which wavelet coefficients to monitor.

Table 10: Comparison of ARLs Under Central Shifts ($\gamma\sigma$)

	γ					
Chart	1.8	2.0	2.2	2.4	2.6	2.8
χ_0^2	6.75	3.89	2.71	1.86	1.49	1.24
T_0^2	5.33	3.26	2.29	1.48	1.43	1.21
T_{B1}^2	3.20	2.15	1.60	1.28	1.13	1.04
T_{B2}^2	5.19	2.94	2.13	1.64	1.45	1.23

3.2.5.2 Shift of a Central Segment

For some signals, data points around the center of a signal are more important in detecting process faults. For example, for the antenna data shown in Figure 1(b) the three main lobes in the center are the most important because they encompass the situations found most frequently in normal usage. Table 10 and Figure 18 give the values of ARL_1. The T_{B1}^2-chart performs slightly better than all other procedures over the entire range of shifts tested. It performs better than the T_0^2-chart when the shifts are less than two σ away from the nominal. The T_{B2}^2-chart performs better than the T_0^2-chart in detecting small vertical shifts, but the T_0^2-chart works better than the T_{B2}^2-chart for larger shifts.

3.2.5.3 Vertical shift

In this case the process mean is shifted vertically (e.g., in Theorem 4, $t_s = t_1$, and $t_e = t_n$ with $\rho_i = \rho,\ i = 1, \ldots, n$). Table 11 and Figure 19 give the resultant ARL_1 values with six different amount of shifts. As expected, the UMPI test-statistic based χ_0^2 chart does not work well for high-dimensional functional data (with $n = 256$), and it has uniformly larger ARL_1 values than other procedures. The procedure based on T_{B1}^2 performs better than the procedure based on T_0^2 in detecting small vertical shifts. For detecting large shifts the conclusion is reversed. Based on Theorem 4 (Remark 3.1), only the wavelet coefficients

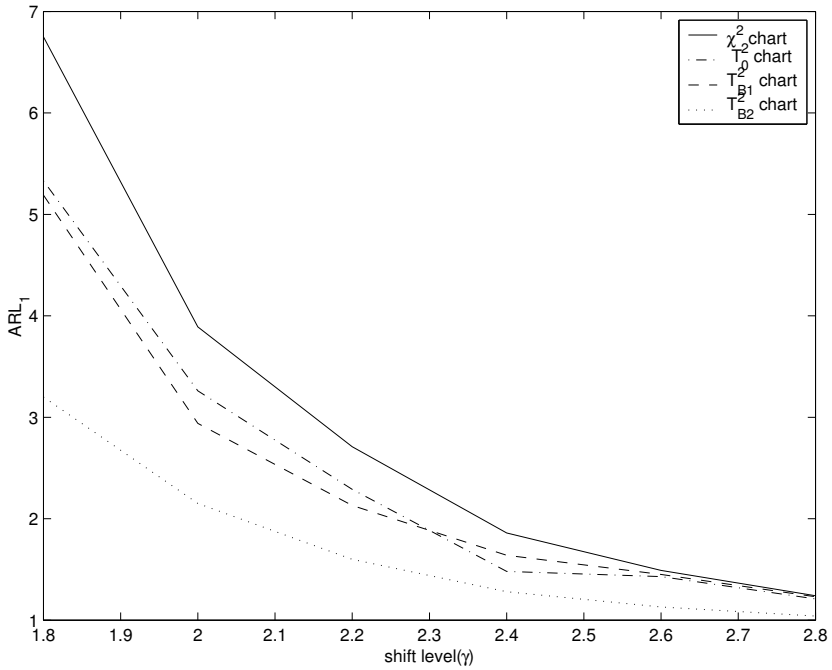

Figure 18: ARL Comparisons Under Central Shifts

Table 11: Comparison of ARLs Under Vertical Shifts ($\gamma\sigma$)

	γ					
Chart	0.1	0.2	0.3	0.4	0.5	0.6
χ^2_0	140.96	50.01	15.23	4.37	1.87	1.21
T^2_0	99.21	17.21	3.65	1.36	1.04	1.00
T^2_{B1}	54.58	27.73	9.64	3.16	1.57	1.08
T^2_{B2}	89.28	35.04	13.18	3.96	1.68	1.13

in the coarser level are affected by the vertical shift. All the wavelet coefficients in the coarser level are always kept for process monitoring in the T^2_0-based SPC charts (Jin and Shi, 2001). It shows good performance for shifts of moderate and large size, e.g., $\gamma = 0.2$ to 0.6. However, for smaller shifts (e.g., $\gamma = 0.1$), the T^2_0-chart compared to the T^2_{B1}-chart shows worse performance because of the noises involved. The T^2_{B2}-chart perfoms similarly to, but slightly worse than, T^2_{B1}-chart.

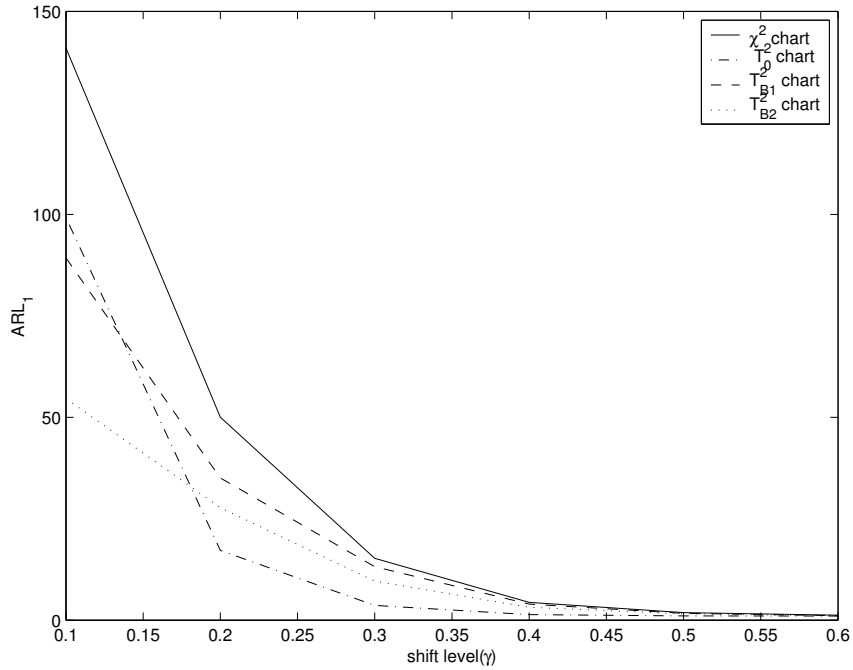

Figure 19: ARL Comparisons Under Vertical Shifts

3.3 An Example Based on Real-life Data Sets

The data set used here was collected at Nortel's production facility in the Research Triangle Park at North Carolina with the goal of developing procedures to detect process problems (Zhou, 1998). The testing equipment receives antenna signals at different degrees of azimuth and elevation. For illustration purposes, we used the central azimuth curve for each of the 18 antenna data sets. Figure 20 (a)-(c) show the runs from nominal processes. Figure 20 (d)-(f) are from faulty processes. Note that they have different patterns of deviation from the nominal processes. Here the dimension (n) of each signal is 256.

Antenna data have numerous "peaks" and "valleys" displaying rather irregular patterns, which present difficulties when modeled by standard statistical procedures. Thus, the our wavelet-based procedure is suitable to handle these data. Follow the robust estimation method used by Donoho and Johnstone (1994). The estimate for the variance of the noise

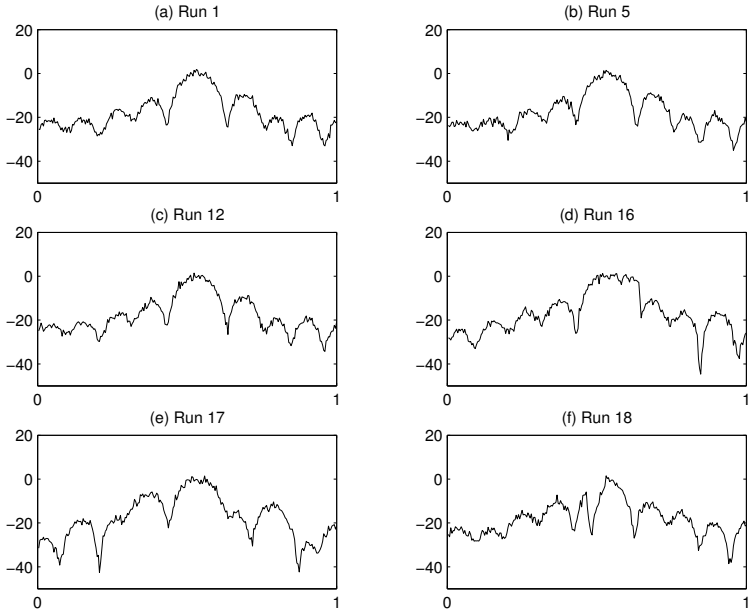

Figure 20: Antenna Data Sets from Different Runs

is obtained as $\hat{\sigma} = \text{median}(|d_{J,k}| : 1 \leq k \leq N/2)/0.6745$ in which J is the finest resolution level (here $J = 8$ from (2)). From the sample curves, the $\hat{\sigma}$ is calculated as 1.0 in our study.

Table 12 displays the values of the test statistic T^2_{B1} for all 18 runs, together with the upper control limit, the optimal threshold value (δ^*), and the size of the selected wavelet-features (m). Using Equation (36), the UCL of T^2_{B1}-chart is calculated based on the type-I error (false-alarm probability) of $\alpha = 0.05$. If the process is in-control, the number of wavelet coefficients to be included in the statistic, T^2_{B1},is small. The statistic T^2_{B1} values are also small. On the other hand, when the process is out-of control, the number of wavelet coefficients included in the statistic is much larger. The T^2_{B1}-statistic value is also much larger. Examination of Table 12 suggests that the first 15 curves are in-control, however, Runs 16, 17, and 18 are out-of control. Note that there are two cases (Run # 3 and # 6) with $T^2_{B1} = 0$. The reason is that the new process data are very close to the nominal process data. This makes the deviates between them small. Thus, no coefficient is selected, i.e., m

Table 12: Results for 18 samples

Run	T^2_{B1}	δ^*	m
1	7.06	2.57	1
2	16.86	2.52	2
3	0.00	0.84	0
4	6.84	2.56	1
5	57.28	2.13	10
6	0.00	2.35	0
7	35.35	2.35	5
8	27.16	2.38	4
9	36.63	2.29	6
10	29.43	2.12	5
11	43.69	2.13	8
12	64.87	2.18	9
13	25.14	2.41	3
14	33.87	2.11	6
15	18.01	2.81	2
16	231.51	3.32	24
17	548.58	3.32	37
18	244.03	3.32	32

$= 0$, and $T^2_{B1} = 0$.

CHAPTER IV

THRESHOLDED SCALOGRAM

4.1 Introduction

Many data signals collected from manufacturing processes are non-stationary and correlated. Many researchers have recommended wavelet-based methods to analyze this type of data (see e.g., Jin and Shi 1999). Wavelet transforms of a signal are multi-resolutional and allow decision-makers to use the information contained in each resolution for signal classification. For example, process fault patterns, which are frequency or phase shifted and invisible to time domain monitoring or control procedures, could be easily detected by wavelet transforms. In particular, Koh *et al.* (1999) indicates that because of its computational efficiency the discrete wavelet transform (DWT) is very useful in on-line (real-time) process monitoring.

One deficiency in the procedures developed from the wavelet coefficients provided from the DWT is the lack of shift-invariance. To elaborate, consider two signals slightly shifted in time. Energy values (e.g., the sum of squared wavelet coefficients) at various frequency scales (or resolution levels) show no difference between the two signals, i.e., the energy is shift-invariant. However, when these signals are transformed and decomposed via the DWT, there is clearly an appreciable difference between the two sets of wavelet coefficients. Therefore, direct assessment of the wavelet coefficients can lead to inaccurate decisions. Thus, a scale-wise energy representation such as a scalogram provides a more robust signal

feature for fault detection against time-shift than the DWT coefficients directly.

Scalograms are commonly used in many fields such as signal and image processing (Rioul and Vetterli 1991) and astronomy and metrology (Scargle 1997). In DWT applications, scalograms measure signal energy contained at various frequency bands with different sizes of scale in wavelet transforms. Intuitively, the scalogram can be useful in monitoring process changes with data collected in time sequences.

In estimating a signal's functional pattern with noisy data, Donoho and Johnstone (1994) proposed a data denoising procedure based on the idea of thresholding out secondary wavelet coefficients representing data noises. In many applications in data mining, the large size of non-stationary data makes computations inefficient (see Pittner and Kamarthi 1999 for an example). Extending the usefulness of the popular scalogram to noisy and possibly massive data, this Chapter develops a thresholded scalogram and studies its properties and applicability to engineering decision making.

4.2 Thresholded Scalograms

Scalograms represent the scale-wise distribution of energies. Scalograms at scale j are defined as (Vidakovic 1999, page 289)

$$S_{dj} = \sum_{k=0}^{2^j-1} d_{jk}^2, j = l, l+1, \ldots, J, \quad \text{and} \quad S_{cl} = \sum_{k=0}^{2^l-1} c_{l,k}^2,$$

where S_{cl} is the energy at the coarsest level. Thus, the total energy of the signal, $||\boldsymbol{y}||^2$, can be decomposed among the resolution levels as follows:

$$||\boldsymbol{y}||^2 = ||\boldsymbol{W}\boldsymbol{y}||^2 = ||\boldsymbol{d}||^2 = \sum_{j=l}^{J} S_{dj} + S_{cl},$$

where $||\boldsymbol{y}||^2 = \sum_{i=1}^{N} y_i^2$. For analyzing potentially massive data and for removing secondary noises, we propose the following thresholded scalogram:

$$S_j^*(\lambda) = \sum_{k=0}^{m_j-1} I(|d_{jk}| > \lambda) d_{jk}^2, \tag{39}$$

where λ is a threshold value that can be selected by various methods (see e.g., Donoho and Johnstone 1994, 1995; Lada *et al.* 2002; Jeong *et al.* 2002) and $m_j = 2^j$ is the number of wavelet coefficients at the jth resolution level. This screening of smaller wavelet coefficients makes the detection of process faults more robust in a noisy environment. Next, we will present the optimal threshold value based on the new criterion that requires balancing data denoising and data reduction goals.

4.2.1 Thresholding Parameter

For a given λ, let $\hat{\boldsymbol{d}}(\lambda) = (\hat{d}_1(\lambda), \ldots, \hat{d}_N(\lambda))^\top$, where $\hat{d}_i(\lambda) = I(|d_i| > \lambda) d_i$, $i = 1, \ldots, N$, be the thresholded wavelet coefficients. Only coefficients larger than λ are kept in subsequent decision making. Later we present the examples that the computation in decision-making (e.g., the decision tree) based on the thresholded scalograms is much more efficient than the use of all original data in the time domain.

In many engineering applications, the relative error,

$$RE = \frac{||\boldsymbol{f} - \hat{\boldsymbol{f}}||}{||\boldsymbol{f}||}, \quad \text{where} \quad \hat{\boldsymbol{f}} = \boldsymbol{W}^{-1}\hat{\boldsymbol{d}}(\lambda),$$

is commonly used in comparing signal approximation quality. See Mallat (1998, page 378-391) for an example employing relative errors in evaluating signal approximation methods. In the equation given below, this type of relative error is used to quantify the modeling accuracy in our data reduction procedure.

Jeong *et al.* (2002) used several real-life data sets and testing curves (e.g., Donoho and Johnstone 1995) to formulate the following objective function for data denoising and data reduction:

$$R_0(\lambda) = \frac{\mathrm{E}(||\boldsymbol{d} - \hat{\boldsymbol{d}}(\lambda)||^2)}{\mathrm{E}(||\boldsymbol{d}||^2)} + \mathrm{E}\left(\frac{||\hat{\boldsymbol{d}}(\lambda)||_0}{N}\right),$$

where $||\hat{\boldsymbol{d}}(\lambda)||_0 = \sum_{i=1}^{N} |\hat{d}_i(\lambda)|_0$, and $|\hat{d}_i(\lambda)|_0 = 1$, if $\hat{d}_i(\lambda) \neq 0$; $|\hat{d}_i(\lambda)|_0 = 0$, otherwise. Note that $||\hat{\boldsymbol{d}}(\lambda)||_0$ is nothing but the number of non-zero $\hat{d}_i(\lambda)$'s. For simplicity, $R_0(\lambda)$ equally weights between the relative error in its first component and the data reduction ratio in its second component. Jeong *et al.* (2002) derived the following theorem for properties of the optimal λ.

Theorem 6 *Consider the model stated in (26). Then,*

(i) the objective function $R_0(\lambda)$ is minimized uniquely at $\lambda = \lambda_R$ where

$$\lambda_R = (\frac{1}{N}\sum_{i=1}^{N}\theta_i^2 + \sigma^2)^{1/2}; \tag{40}$$

The moment estimate of λ_R,

$$\hat{\lambda}_R = \left(\frac{1}{N}\sum_{i=1}^{N} d_i^2\right)^{1/2}, \tag{41}$$

has the following properties:

(ii) $(\hat{\lambda}_R - \lambda_R)$ converges to 0 with probability one;

(iii) $\sqrt{N}(\hat{\lambda}_R - \lambda_R)/\sigma_N^*$ converges in distribution to $N(0,1)$, where

$$(\sigma_N^*)^2 = \frac{1}{4N}\left(\frac{4\sigma^2\sum_{i=1}^{N}\theta_i^2 + 2N\sigma^4}{\sum_{i=1}^{N}\theta_i^2 + \sigma^2}\right).$$

Remark: The λ in (39) could be replaced by any estimate such as $\hat{\lambda}_R$ from the above method or $\hat{\lambda}_D$ from Donoho and Johnstone (1995). When the estimate converges to its λ_R or λ_D with probability one, the following lemmas and theorems will carry through. Thus, for the remainder of this article, we will use λ_N and $\hat{\lambda}_N$ to represent these parameters and estimates.

4.3 Asymptotic Properties of Thresholded Scalograms

Replacing λ by $\hat{\lambda}_N$ in (39), we obtain the following applicable thresholded scalogram:

$$\hat{S}^*_{Nj}(\hat{\lambda}_N) = \sum_{k=0}^{m_j-1} I(|d_{jk}| > \hat{\lambda}_N) d^2_{jk}.$$

For convenience of notation, let

$$S^*_{Nj} = \sum_{k=0}^{m_j-1} I(|d_{jk}| > \lambda_N) d^2_{jk}, \tag{42}$$

which represents the thresholded scalogram with respect to a thresholding parameter λ_N.

Note that d_{jk}'s are independent, but are not identically distributed because of different means. First, we will use the following lemma to show that the difference between S^*_{Nj} and $\hat{S}^*_{Nj}$ converges to zero with probability one. Then, we will focus on the derivation of the asymptotic distribution of S^*_{Nj}.

Lemma 1 *Assume that* $\{d_i : i = 1, 2, \dots\}$ *is a series of independent random variables with possibly different means and the same variance. For a fixed index* i*, let* $\hat{Y}_{Ni} = I(|d_i| \geq \hat{\lambda}_N)$ *and* $Y_{Ni} = I(|d_i| \geq \lambda_N)$*, where* $\hat{\lambda}_N$ *and* λ_N *are defined in (40) and (41), respectively. Then*

$$\hat{Y}_{Ni} - Y_{Ni} \xrightarrow{w.p.1} 0 \quad as \quad N \to \infty.$$

Proof: For any $\varepsilon > 0$,

$$\begin{aligned}
& \Pr\{\lim_{N\to\infty} |\hat{Y}_{Ni} - Y_{Ni}| > \varepsilon\} \\
= \; & \Pr\{\lim_{N\to\infty} (\hat{Y}_{Ni} = 1, Y_{Ni} = 0)\} + \Pr\{\lim_{N\to\infty} (\hat{Y}_{Ni} = 0, Y_{Ni} = 1)\} \\
= \; & \Pr\{\lim_{N\to\infty} (|d_i| \geq \hat{\lambda}_N, |d_i| < \lambda_N)\} + \Pr\{\lim_{N\to\infty} (|d_i| < \hat{\lambda}_N, |d_i| \geq \lambda_N)\} \\
= \; & \Pr\{\lim_{N\to\infty} (\hat{\lambda}_N \leq |d_i| < \lambda_N)\} + \Pr\{\lim_{N\to\infty} (\lambda_N \leq |d_i| < \hat{\lambda}_N)\} \\
\leq \; & \Pr\{\lim_{N\to\infty} (\hat{\lambda}_N < \lambda_N)\} + \Pr\{\lim_{N\to\infty} (\lambda_N < \hat{\lambda}_N)\} \\
= \; & \Pr\{\lim_{N\to\infty} (\hat{\lambda}_N - \lambda_N \neq 0)\} \\
= \; & 0.
\end{aligned}$$

The last equation follows from $(\hat{\lambda}_N - \lambda_N) \xrightarrow{w.p.1} 0$.

□

Based on this lemma, the corresponding result for the thresholded scalogram can be obtained immediately. Theorem 7 states this result without proof.

Theorem 7 *Under the conditions in Theorem 6, for fixed j, $j = l, l+1, \ldots, J$, we have*

$$(\hat{S}^*_{Nj} - S^*_{Nj}) \xrightarrow{w.p.1} 0 \;\; as \; N \to \infty.$$

Next, we derive the asymptotic distribution of S^*_{Nj}. Recall that d_{jk}'s in S^*_{Nj} are independent and normally distributed with mean θ_{jk} and common variance σ^2. Let ϕ as the

probability density function of a standard normal distribution. Then

$$\begin{aligned}
\mathrm{E}(S^*_{Nj}) &= \sum_{k=0}^{m_j-1} \mathrm{E}(I(|d_{jk}| \geq \lambda_N) d^2_{jk}) = \sum_{k=0}^{m_j-1} \int_{|t|\geq\lambda_N} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt \\
&= \sum_{k=0}^{m_j-1} (\theta^2_{jk} + \sigma^2) - \sum_{k=0}^{m_j-1} \int_{|t|<\lambda_N} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(S^*_{Nj}) &= \sum_{k=0}^{m_j-1} \mathrm{Var}(I(|d_{jk}| \geq \lambda_N) d^2_{jk}) \\
&= \sum_{k=0}^{m_j-1} \left(\mathrm{E}(I(|d_{jk}| \geq \lambda_N) d^4_{jk}) - \mathrm{E}^2(I(|d_{jk}| \geq \lambda_N) d^2_{jk})\right) \\
&= \sum_{k=0}^{m_j-1} \left[\mathrm{E}(d^4_{jk}) - \int_{|t|<\lambda_N} t^4 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt\right] \\
&\quad - \sum_{k=0}^{m_j-1} \left[\theta^2_{jk} + \sigma^2 - \int_{|t|\leq\lambda_N} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt\right]^2 \\
&= \sum_{k=0}^{m_j-1} \left[(3\sigma^4 + 6\sigma^2\theta^2_{jk} + \theta^4_{jk}) - \int_{|t|<\lambda_N} t^4 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt\right] - \\
&\quad \sum_{k=0}^{m_j-1} \left[\theta^2_{jk} + \sigma^2 - \int_{|t|\leq\lambda_N} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt\right]^2 .
\end{aligned}$$

These support the proof of the following theorem.

Theorem 8 *Let* $\eta_j = \mathrm{E}(S^*_{Nj}) \geq 0$ *and assume that* $Var(S^*_{Nj})/m_j \to \sigma^2_j$ *as* $m_j \to \infty$. *Then, under the conditions in Theorem 1*

$$\frac{\eta_j(\ln S^*_{Nj} - \ln \eta_j)}{\sqrt{m_j}\ \sigma_j} \xrightarrow{d} N(0,1) \quad as \quad m_j \to \infty.$$

Proof: Let $X_{jk} = d^2_{jk} I(|d_{jk}| > \lambda_N)$. These X_{jk}'s are independent random variables with the finite mean $\mathrm{E}(X_{jk}) = \mu_{jk}$ and the finite variance $\mathrm{Var}(X_{jk}) = \sigma^2_{jk}$. Then, $\eta_j =$

$\mathrm{E}(S^*_{Nj}) = \sum_{k=0}^{m_j-1} \mu_{jk}$ and $\mathrm{Var}(S^*_{Nj}) = \sum_{k=0}^{m_j-1} \sigma^2_{jk}$. To show the asymptotic normality of $(S^*_{Nj} - \eta_j)/(\sqrt{m_j}\sigma_j)$, it is sufficient to verify the following Lindeberg condition (Serfling 1980, page 30), for each fixed $\varepsilon > 0$,

$$\frac{1}{m_j} \sum_{k=0}^{m_j-1} \int_{|t^2 I(|t|>\lambda_N)-\mu_{jk}|>\varepsilon\sqrt{m_j}} [t^2 I(|t| > \lambda_N) - \mu_{jk}]^2 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt \to 0 \ \ as \ \ m_j \to \infty.$$

It follows that

$$\begin{aligned}
& \int_{|t^2 I(|t|>\lambda_N)-\mu_{jk}|>\varepsilon\sqrt{m_j}} [t^2 I(|t| > \lambda_N) - \mu_{jk}]^2 \phi\left(\frac{(t-\theta_{jk})}{\sigma}\right) dt \\
= \ & O\left(\int_{t^2>\varepsilon\sqrt{m_j}} t^4 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt\right) \\
= \ & O\left(\int_{t>\varepsilon^{1/2} {m_j}^{1/4}} t^4 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt\right) \\
= \ & O\left(\varepsilon^2 m_j \phi\left(\frac{\varepsilon^{1/2} {m_j}^{1/4} - \theta_{jk}}{\sigma}\right)\right) \\
= \ & O\left(\varepsilon^2 m_j \exp\left\{-\frac{\varepsilon\sqrt{m_j}}{2\sigma^2}\right\}\right).
\end{aligned}$$

Therefore, for every $\varepsilon > 0$, as $m_j \to \infty$,

$$\int_{|t^2 I(|t|>\lambda_N)-\mu_{jk}|>\varepsilon\sqrt{m_j}} [t^2 I(|t| > \lambda_N) - \mu_{jk}]^2 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) dt = O\left(\varepsilon^2 m_j \exp\left\{-\frac{\varepsilon\sqrt{m_j}}{2\sigma^2}\right\}\right) \to 0.$$

Recall the delta method: if $(T_N - \eta_N)/\tau_N \xrightarrow{d} N(0,1)$, then $[h(T_N) - h(\eta_N)]/[\tau_N h'(\eta_N)] \xrightarrow{d} N(0,1)$ provided h is a continuous function such that $h'(\eta_N)$ exists and $h'(\eta_N) \neq 0$. By applying the delta method with $h(\eta_N) = \ln \eta_N$ and $h'(\eta_N) = 1/\eta_N$, we obtain the stated result.

□

Corollary 1 *Let* $\eta^*_{Nj} = \mathrm{E}(\hat{S}^*_{Nj}) \geq 0$, *and assume that* $Var(\hat{S}^*_{Nj})/m_j \to (\sigma^*_{Nj})^2$ *as* $m_j \to \infty$.

Then, under the conditions in Theorem 1

$$\frac{\eta^*_{Nj}(\ln \hat{S}^*_{Nj} - \ln \eta^*_{Nj})}{\sqrt{m_j}\ \sigma^*_{Nj}} \xrightarrow{d} N(0,1)$$

as $N \to \infty$ *and* $m_j \to \infty$.

4.4 Application of Scalograms for Fault Detection and Classification

4.4.1 Fault Detection Using Thresholded Scalograms

The asymptotic distribution of the thresholded scalograms can be used to establish an approximate $100(1-\alpha)\%$ confidence interval, $\ln \hat{S}^*_{Nj} \pm z_{\alpha/2}\hat{\sigma}^*_{Nj}/(\hat{\eta}^*_{Nj})$, where z_α is the upper $100(1-\alpha)\%$ percentile of the standard normal distribution. By connecting the point-wise interval values for each resolution level as shown in Figure 21, we can construct a set of lower and upper bounds of thresholded scalograms for the nominal run. This will serve as a tool to statistically detect process faults at several resolution levels. This idea is applied to a rapid thermal chemical vapor deposition (RTCVD) process that deposits thin films on semiconductor wafers using a temperature-driven surface chemical reaction.

As feature size decreases, the functional operation of devices (e.g., transistors) becomes increasingly susceptible to failure because of variations in deposition processes. Therefore, detecting a process condition different from the nominal is critical.

Quadruple mass spectrometry (QMS) is commonly used in the semiconductor manufacturing processes for monitoring thin-film deposition quality. Figure 22 shows the control system to predict the volume of silicon deposited from the in situ QMS sensor data (Smith, 1998). By monitoring the carrier gas (Ar^+) signal during a process run, we can control the thickness of deposited film. Figure 23 presents data collected by the QMS in a research

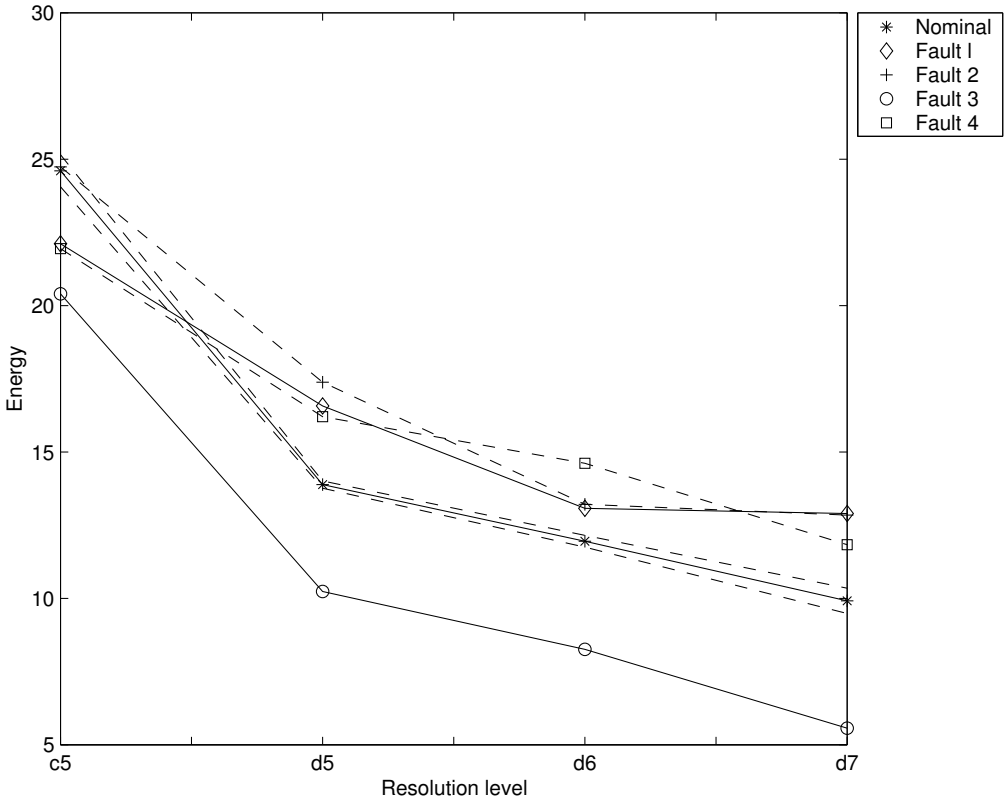

Figure 21: Point-wise Confidence Intervals of Thresholded Scalograms for the Nominal Run.

project (Rying, 1997) to develop an in-situ measurement technique for online process monitoring. The subfigures represent one of the 21 nominal RTCVD process runs and four sets of data from different faulty processes. Although only 128 data points are in the curve and the data change pattern is not very complicated, this case study serves as a basis for developing process monitoring and fault detection/classification tools applicable to many engineering applications. More important, wavelet transforms have proven to be useful in locating those change points for the in-situ deposition thickness measurement tool (Lada *et al.* 2002) for thin films.

Comparably, the scalogram values for the data in the Fault 3 class are much different from the nominal one at all resolution levels. Because of the similarity of the data curves

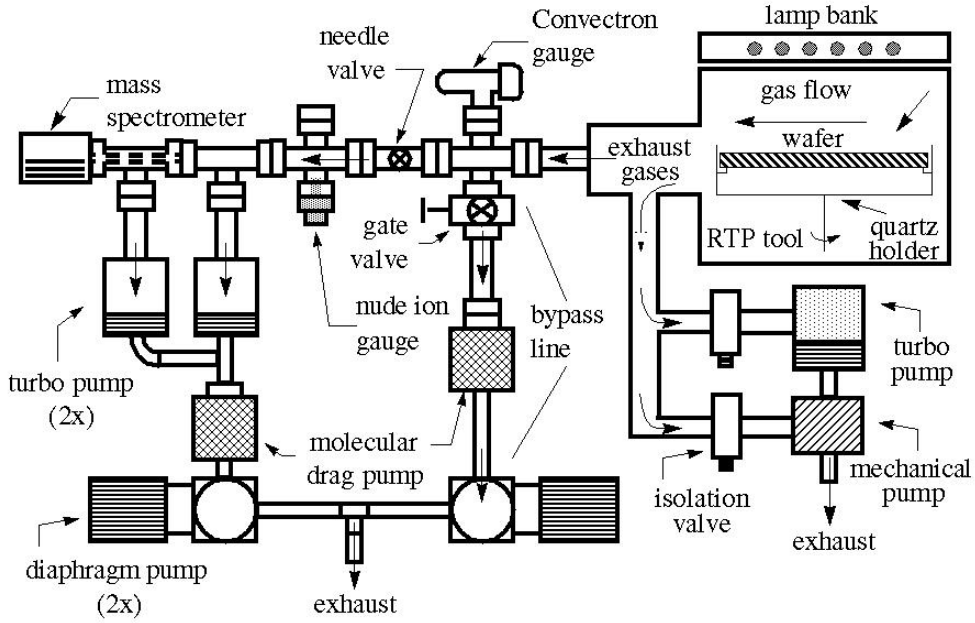

Figure 22: Schematic of the QMS Sensor Apparatus and Adjoining RTP Tool

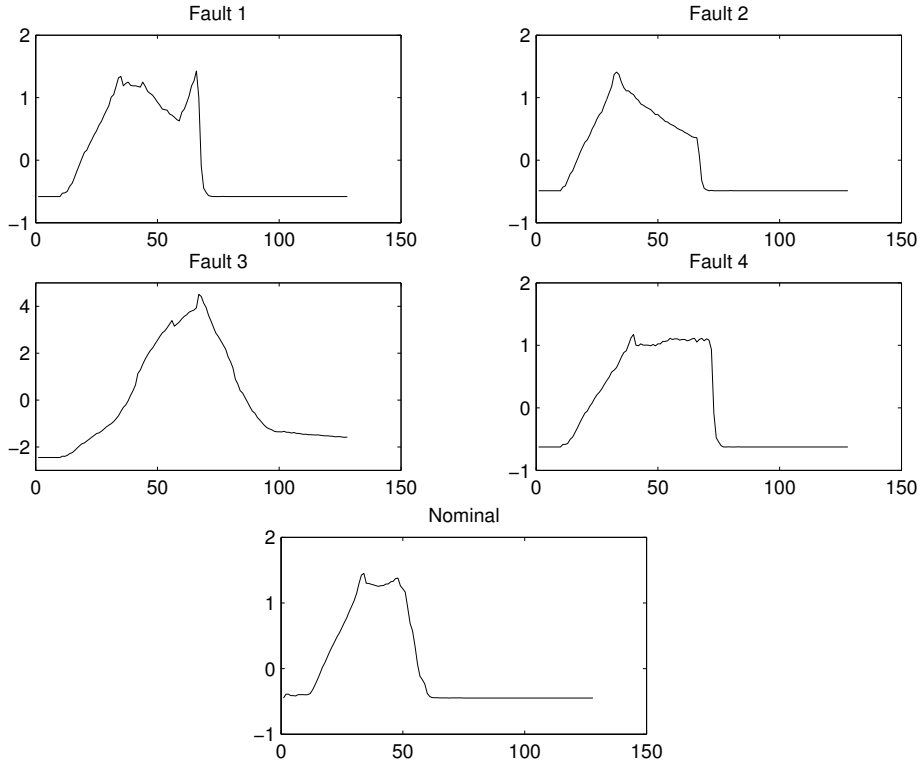

Figure 23: RTCVD Signals.

in the original time domain, Fault classes 1 and 2 have similar scalogram values at the finer resolution levels, but not at the coarsest resolution level. The sharp drop in the curve in the Fault Class 1 may partly explain why the value of its coarsest level scalogram is much different from its nominal value as compared with the value obtained from the Fault Class 2. The Fault Class 2 and the nominal curves have similar finer and coarsest level scalogram values, but this is not seen at the middle level of the scalograms. Results plotted in Figure 21 show that these four classes of curves are clearly out of bounds at almost all resolution levels except at the coarsest level for Fault 2 class.

4.4.2 Data Mining Using Thresholded Scalograms

This subsection presents another example with a testing curve (Mallat 1998, page 378). Figure 24 shows the curves from the nominal run (the original signal pattern as given in Mallat (1998)) and three fault-situations artificially created for experimenting with the applicability and sensitivity of the proposed metric. Note that these testing curves have many sharp peaks and drops that are difficult for most statistical techniques to detect.

Figure 25 shows the results of a clustering analysis based on thresholded scalograms. Thresholded scalograms at the fifth and sixth levels were used as features for clustering. This plot shows that these four signals can be well discriminated based on thresholded scalograms. This example illustrates the potential of the scalograms for signal classification.

Next, we apply a commonly used data mining tool CART (Classification and Regression Trees) to the thresholded scalograms for analyzing these signals. See Breiman *et al.* (1984) for details of CART tree-building and pruning procedures.

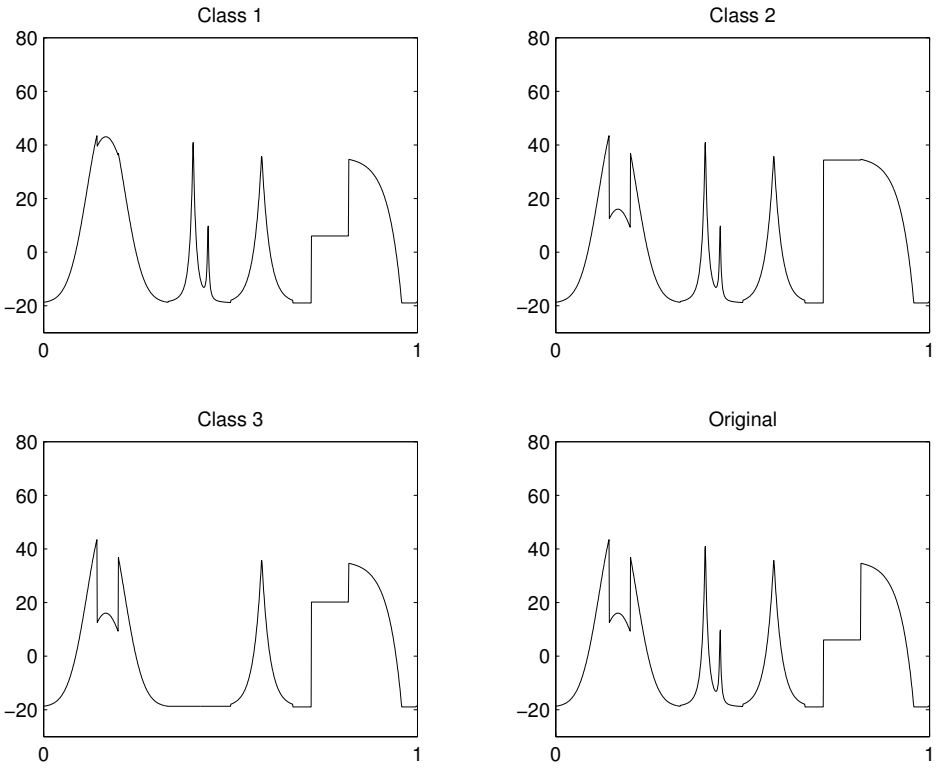

Figure 24: Four Classes of Piecewise Signals

For applying CART to the four classes of curves presented in Figure 24, various curve-replicates were generated. In our experiment, all of the testing curves were shifted to the left (or right) in 5 (or 10, 15, 20, 25, 30) time-units (out of a total of $N = 1,024$ units) for generating a new curve. Moreover, Gaussian random noises with $\sigma = 0.1$ are also added. Shifting the curves to the left and right artificially tested the invariance property of thresholded scalograms. For all curves in these four classes, the above data replication method was applied in order to generate 1,200 total replicated-curves (300 in each case). CART will then identify all these fault types based on the scalogram data.

Some of the curves from fault conditions are considerably more difficult for decision trees to correctly identify. For example, the only difference between class 1 and the original curve is a smaller vertical drop in the first rectangle-shape dip located around 147 to 204 time

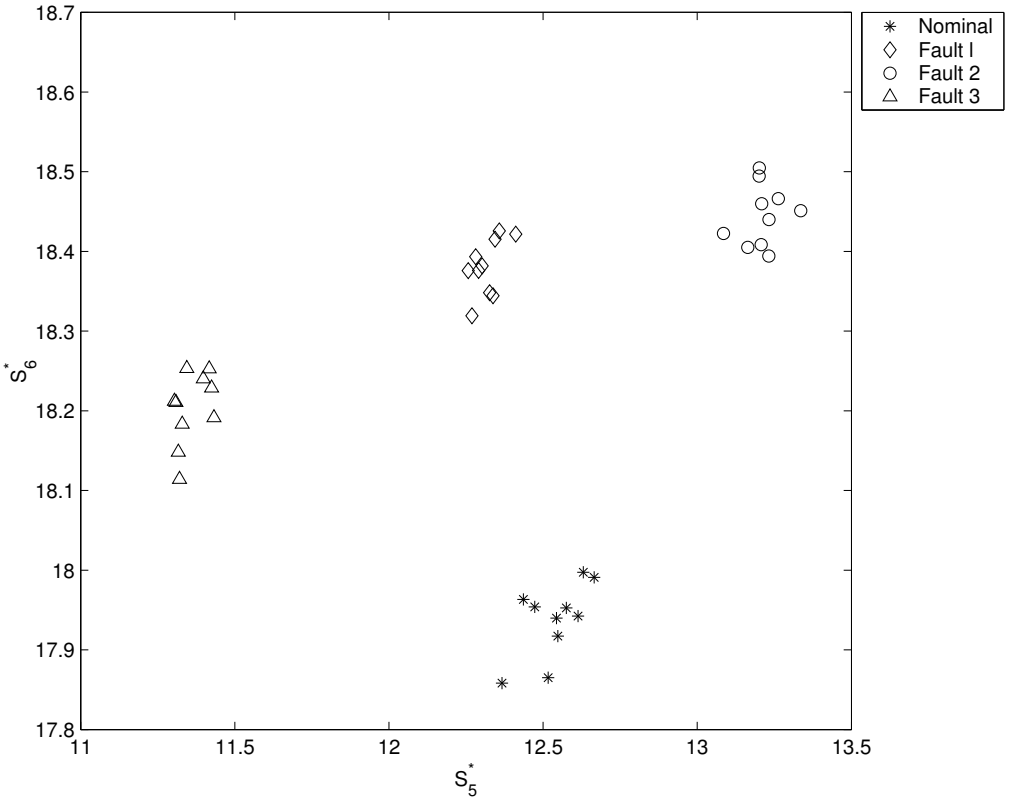

Figure 25: Clustering Using Thresholded Scalograms

units. In Figure 26, the notation S^*_{c5} represents the at the coarsest level and S^*_{dj} the energy at the finer resolution level j. The first split is $S^*_{d9} \leq 6.54$ where S^*_{d9} is the energy at the finest resolution level. If $S^*_{d9} > 6.54$, then the signal is assigned to class 2; otherwise, one goes to node 2. Similar interpretations could be obtained for other nodes.

Table 13 shows the importance-rankings of variables selected from CART. The scores reflect the contribution each variable made in classifying or predicting the target variable; in each case the contribution stems from the role of each variable as both a primary splitter and as a surrogate to any of the primary splitters. The relative importance of input variables can be measured by

$$\hat{I}_j = \sum_{t=1}^{J-1} \hat{i}_t^2 \; I(v_t = j),$$

where the summation is over all the non-terminal nodes t of the J-terminal node tree, v_t

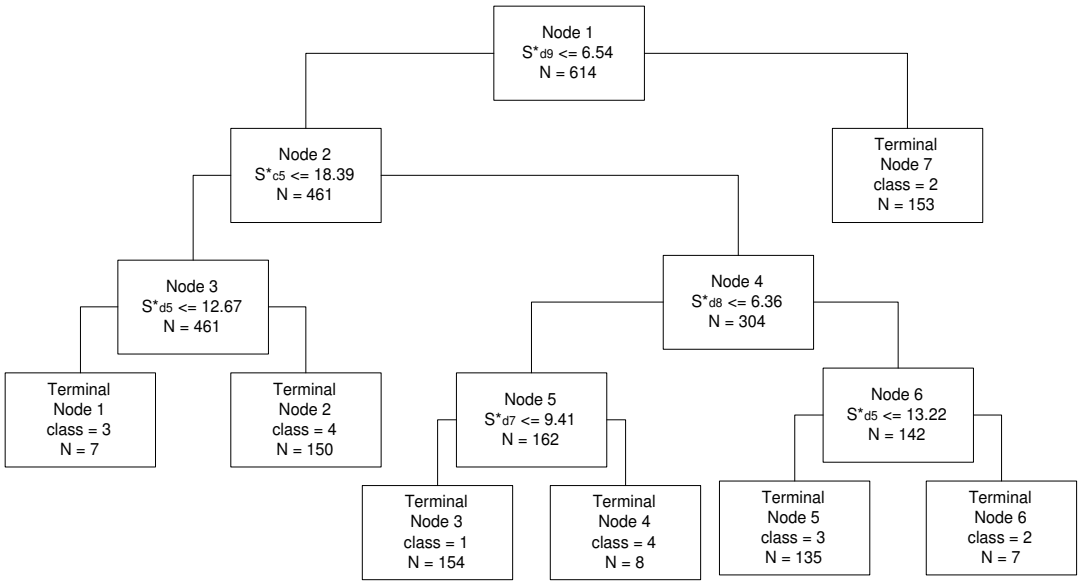

Figure 26: CART Tree Using Thresholded Scalograms

Table 13: Variable Importance

Variable	S^*_{d5}	S^*_{d6}	S^*_{d8}	S^*_{c5}	S^*_{d9}	S^*_{d7}
Score	100.00	91.32	70.42	69.57	5.68	55.19

is the splitting variable associated with node t, and $\hat{i}_t^2$ is the empirical improvement in misclassification error as a result of a split (Breiman *et al.* 1984). The influence of the most influential variable is arbitrarily assigned the value 100. In our example, S^*_{d5} is ranked most important. Note that the thresholded scalograms, S^*_{d5} and S^*_{d6}, at two finer resolution levels (but not at the coarsest level S^*_{c5}) are more important than the others in this example.

The misclassification rates in the training and testing samples are shown in Table 14. All classification errors are less than 4%, with many error-free cases. Our scalogram-based CART method performed well despite complicated testing curves, noises, and left-and-right shifting that could have made classification difficult.

Table 14: Misclassification error(%)

Class	Training data	Testing data
original	0.65	0.69
1	1.91	0.00
2	0.00	0.00
3	0.70	3.82

CHAPTER V

CONCLUSIONS AND FUTURE RESEARCH

5.1 Summary of Results

5.1.1 Wavelet-Based Data Reduction Procedures

In this research, we proposed an idea for handling a special type of large size nonstationary data in data analysis and decision making. The properties of data reduction methods are investigated by testing two real-life examples and the many popular data signals in the literature of statistics and engineering. Several evaluation studies with popular testing curves used in the literature and with two real-life data sets demonstrate the superiority of the proposed methods over engineering data compression and statistical data de-noising methods that are currently used to achieve data reduction goals. Results from the classification trees show that the proposed methods are at least as accurate, and sometimes more so, as the results obtained from analysis of the original larger size data; however, the proposed methods have a clear advantage in computational efficiency.

5.1.2 SPC Procedures for Nonstationary Functional Data

We presented several statistical process control charting (SPC) procedures for functional data by utilizing the special ability of the discrete wavelet transform to model sharp-change data. Based on the simulation studies, the T^2_{B1}-chart is generally more effective for detecting many kinds of process changes, whether globally or locally, than the method represented in the UMPI χ^2-chart and other charts extended from ideas given in the literature. In all

cases, the proposed methods are considerably more effective in detecting smaller shifts. Our procedure worked well given the lack of any prior information on which wavelet coefficients to monitor (i.e., what type of process changes at what location of the process data).

5.1.3 Thresholded Scalogram and Its Applications in Process Fault Detection

In this research, we proposed the use of thresholded scalograms to detect process faults in processes with noisy and possibly massive data that exhibit time-shifted patterns. The properties of thresholded scalograms were explored via theoretical and empirical investigations. One real-life example and one simulated case study were presented to illustrate the potential of the proposed method. We believe that when large amounts of data are involved, our procedure will become even more powerful and important.

5.2 Future Research

Future work is needed to explore the strengths and weaknesses in other rules used for decisions (e.g., clustering analysis in data mining) and to extend the proposed idea to traditional quality improvement and SPC areas (e.g., to analyze the design of experiment data based on reduced-size information, analysis of the variance of time-sequence or spatial data based on thresholded wavelet coefficients, and multi-resolution SPC for spatial image data in process monitoring). We will also consider extending the above to highly dimensional data, e.g. imagery data set. Newly developed multiscale methods for high-dimensional data, such as beamlets, wedgelets (Donoho and Huo 2001), and so on will be explored.

Since the exponentially weighted moving average (EWMA) and CUSUM procedures are so popular in the SPC literature, extensions of our methods to these types of control charts are important. Phase-I studies for establishing process parameters are needed. Note that

the research of wavelet-thresholding procedures for multiple data-curves representing the baseline process is very limited. Further research in this area is needed. Extension from monitoring mean changes to variance changes is needed to handle problems encountered in the studies such as Ganesan *et al.* (2002). Our proposed SPC methods need to be extended to a general covariance structure for both cases of known and unknown covariance matrices. The choice of both wavelets and decomposition level can affect the ARL performance, and we need research to determine the appropriate wavelet type and decomposition level. We need to develop a new generation of SPC procedures for both monitoring and classifying the root causes of process problems.

REFERENCES

[1] Alt, F.B. (1985), "Multivariate Quality Control," in *Encyclopedia of Statistical Sciences 6*, edited by S. Kotz and N. L. Johnson, John Wiley & Sons, New York, NY.

[2] Antoniadis A., Gijbels, I., and Grégoire, G. (1997), "Model S election Using Wavelet Decomposition and Applications," *Biometrika*, 84(4), 751–763.

[3] Bakshi, B. R. (1999), "Multiscale Analysis and Modeling using Wavelets," *Journal of Chemometrics*, 13, 415–434.

[4] Braha, D. (2001), *Data Mining for Design and Manufacturing: Methods and Applications*, Kluwer Academic Publishers.

[5] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, New York: Chapman and Hall.

[6] Bruce, A. G., and Gao, H.-Y. (1996), "Understanding WaveShrink: Variance and Bias Estimation," *Biometrika*, 83(4), 727–745.

[7] Daubechies, I. (1992), *Ten Lectures in Wavelets*, John Wiley, Philadelpia.

[8] Doganaksoy, N., Faltin, F. W., and Tucker, W. T. (1991), "Identification of Out-of-Control Quality Characteristics in a Multivariate Manufacturing Environment," *Communications in Statistics- Theory and Methods*, 20, 2775-2790.

[9] Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81(4), 425–455.

[10] Donoho, D. L., and Johnstone, I. M. (1995), "Adapting to Unknown Smoothness via Wavelet Ahrinkage," *Journal of the American Statistical Association*, 90(432), 1200–1224.

[11] Donoho, D. L., and Huo, X. (2001), "Beamlets and Multiscale Image Analysis," in *M*ultiscale and Multiresolution Methods, Editors T.J. Barth, T. Chan, and R. Haimes, Springer Lecture Notes in Computational Science and Engineering, 20, 149–196.

[12] Fan, J. (1996), "Test of Significance Based on Wavelet Thresholding and Neyman's Truncation," *J. American Statistical Association*, 91, 674-688.

[13] Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, London: Morgan Kaufmann Inc.

[14] Forsythe, G.E., Malcolm, M.A., and Moler, C.B. (1976), *Computer Methods for Mathematical Computations*, Prentice Hall.

[15] Ganesan, R., and Das, T. K., (2002), "Wavelet Based Multiscale Statistical Process Monitoring-Literature Review and Research Extensions", submitted to *IIE Trans. on Quality and Reliability*

[16] Ganesan, R., Das, T. K., Sikder, A. K., and Kumar, A. (2002), "Wavelet Based Indentification of Delamination Emission Signal", submitted to *IEEE Trans. on Semiconductor Manufacturing*

[17] Gardner, M. M., Lu, J. C., Gyurcsik, R. S., Wortman, J. J., Hornung, B. E., Heinisch, H. H., Rying, E. A., Rao, S., Davis, J. C., and Mozumder, P. K. (1997), "Equipment Fault Detection Using Spatial Signatures," *IEEE Transaction on Components,*

Packaging, and Manufacturing Technology-Part C, Vol. 20, pp.295–303.

[18] Gnanadesikan, R. and kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics*, 28, 81–124.

[19] Hall P., and Poskitt, D. S. (2001), "A Functional Data-Analytic Approach to Signal Discrimination," *Technometrics*, 43(1), 1–9.

[20] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning* , Springer-Verlag.

[21] Hawkins, D. M. (1991), "Multivariate Quality Control Based on Regression-Adjusted Variables," *Technometrics*, 33, 61–76.

[22] Hawkins, D. M. (1993), "Regression Adjustment for Variables in Multivariate Quality Control," *Journal of Quality Technology*, 25, 170–182.

[23] Hotelling, H. (1947), "Multivariate Quality Control," *Techniques of Statistical Analysis* (Eisenhart, C., Hastay, M., and Wallis, W. A. eds.), McGraw-Hill, New York, NY, 111-184.

[24] Ihara, I. (1993), *Information Theory for Continuous System,* New Jersey: World Scientific.

[25] Jackson, J. E. (1980), "Principal Components and Factor Analysis: Part I- Principal Components," *Journal of Quality Technology*, 12, 201–213.

[26] Jeong, M. K., Lu, J. C., Huo, X., Vidakovic, B., and Chen, D. (2002), "Wavelet-Based Data Reduction Techniques for Fault Detection and Classification," *submitted to Technometrics*.

[27] Jeong, M. K., Chen, D. and Lu, J. C. (2003), "Thresholded Scalogram and Its Application in Process Fault Detection," *Applied Stochastic Models in Business and Industry*, 19(3), 231-244.

[28] Jin, J., and Shi, J. (1999), "Feature-Preserving Data Compression of Stamping Tonnage Information Using Wavelets," *Technometrics*, 41(4), 327–339.

[29] Jin, J., and Shi, J. (2001), "Automatic Feature Extraction of Waveform Signals for In-Process Diagnostic Performance Improvement," *Journal of Intelligent Manufacturing*, 12, 257-268.

[30] Jones, M. C., and Rice, J. A. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *American Statistician*, 46, 140-145.

[31] Jung, U., and Lu, J. C. (2004), "A Wavelet-based Random-effect Model for Multiple Sets of Complicated Functional Data," *Technical Report*, The School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA. See http://www.isye.gatech.edu/the paper.

[32] Kang, L., and Albin, S. L. (2000), "On-Line Monitoring When the Process Yields a Linear Profile," *Journal of Quality Technology*, 32, 418-426.

[33] Kasashima, N., Mori, K., Ruiz., G. H. and Taniguchi, N. (1995), "On-Line Failure Detection in Face Milling Using Discrete Wavelet Transform," *Annals of the CIRP*, 44, 483-487.

[34] Kiefer, J., and Schwartz, R. (1965), "Admissible Bayes Character of T^2-, R^2-, and Other Fully Invariant Tests for Classical Multivariate Normal Problems," *Ann. Math. Statist.*, 36, 747-770.

[35] Kim, K., Mahmoud, M. A., and Woodall, W. H. (2003), "On the Monitoring of Linear Profiles," To appear in *Journal of Quality Technology*.

[36] Koh, C. K. H., Shi, J., Williams, W. J., and Ni, J. (1999), "Multiple Fault Detection and Isolation Using the Haar Transform, Part 2: Application to the Stamping Process," *Transactions of the ASME*, 295–299.

[37] Kudo, A. (1963), "A Multivariate Analogue of the One-Sided Test," *Biometrika*, 50, 403-418.

[38] Lada, E. K., Lu, J. C., and Wilson, J. R. (2002), "A Wavelet Based Procedure for Process Fault Detection," *IEEE Trans. on Semiconductor Manufacturing*, 15(1), 79–90.

[39] Lawless, J. F., Mackay, R. J., and Robinson, J. A. (1999), "Analysis of Variation Transmission in Manufacturing Process-Part ," *Journal of Quality Technology*, 31, 131–142.

[40] Liu, B., and Ling, S. F. (1999), "On the Selection of Informative Wavelets for Machinery Diagnosis," *Mechanical Systems and Signal Processing*, 13(1), 145–162.

[41] Mallat, S. G. (1998), *A Wavelet Tour of Signal Processing*, San Diego: Academic Press.

[42] Mallet, Y., Coomans, D., Kautsky, J., and De Vel, O. (1997), "Classification Using Adaptive Wavelets for Feature Extraction," IEEE Transaction on Pattern Analysis and Machine Intelligence, 19(10), 1058–1066.

[43] Martell, L. (2000), "Data Reduction and Model Selection with Wavelet Transforms," unpublished Ph.D. thesis, Department of Statistics, North Carolina State University, Raleigh, North Carolina.

[44] Mahmoud A. M. and Woodall, W. H. (2002), "Phase I Monitoring of Linear Profiles with Calibration Applications," *submitted to Technometrics*.

[45] Mallat, S. G. (1989), "A Theory for Multiresolution Signal Decomposition: the Wavelet Representation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 11, pp.674-693, October.

[46] Montgomery, D. C. (2001), *Introduction to Statistical Quality Conrol*, 4th Edition, Hohn Wiley & Sons, New York, NY.

[47] Morettin, P. (1997), "Wavelets in Statistics," *Resenhas*, 3(2), 211-272.

[48] Mori, K., Kasashima, N., Yoshioka, T. and Ueno, Y. (1996), "Prediction of Spalling on a Ball Bearing by Applying the Discrete Wavelet Transform to Vibration Signals," *Wear 195, Elsevier Sciences S. A.*, pp. 162-168.

[49] Nair, V. N., Taam, W., and Ye, K. Q. (2002), "Analysis of Functional Responses from Robust Design Studies," *Journal of Quality Technology*, 34, 355-370.

[50] Perlman, M. D. (1969), "One-Sided Testing Problems in Multivariate Analysis," *Ann. Math. Statist.*, 40, 549-567.

[51] Pittner, S., and Kamarthi, V. (1999), "Feature Extraction From Wavelet Coefficients for Pattern Recognition Tasks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(1), 83–88.

[52] Ramsay, J. O., and Silverman, B. W. (1997). *Functional Data Analysis.* New York: Springer-Verlag.

[53] Rao, R. M., and Bopardikar, A. S. (1998), *Wavelet Transforms: Introduction to Theory and Applications*, Reading, Massachusetts: Addison-Wesley.

[54] Rencher, A. C. (1993), "The Contribution of Individual Variables to Hotelling's T^2, Wilk's Λ, and R^2," *Biometrics*, 49, 479–489.

[55] Rioul, O., and Vetterli, M. (1991), "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, October, 14–38.

[56] Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, the University Press, Cambridge.

[57] Roy, J. (1958), "Step Down Procedure in Multivariate Analysis," *Ann. Math. Statist.*, 29, 1177-1187.

[58] Rying, E. A. (2001), "A Novel Focused Local-learning Wavelet Network with Application to In-situ Selectivity and Thickness Monitoring During Selective Silicon Epitaxy," Ph.D. Thesis, Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, Carolina.

[59] Saito, N. (1994), "Simultaneous Noise Suppression and Signal Compression Using a Library of Orthonormal Bases and the Minimum Description Length Criterion," in

Wavelets in Geophysics, E. Foufoula-Georgiou and P. Kumar, Eds. New York: Academic Press, 299–324.

[60] Scargle, J.D. (1997), "Wavelet Methods in Astronomical Time Series Analysis," in *Application of time seires analysis in astronomy and meteorology*, T. S. Rao, M. B. Priestly, and O. Lessi, Eds. New York: Chapman and Hall, 226–248.

[61] Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: An Introduction with Applications*, London: Cambridge University Press.

[62] Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

[63] Silvapulle, M. J. (1995), "A Hotelling's T^2-type Statistic for Testing Against One-Sided Hypothesis," *J. Multivariate Analysis*, 55, 312–319.

[64] Smith, R. M. (1998), "Real Time Control of Polysilicon Deposition in Single-Wafer Rapic Thermal Chemical Vapor Deposition Furnaces," unpublished Ph.D. thesis, Department of Statistics, North Carolina State University, Raleigh, North Carolina.

[65] Srivastava, M. S., and Worsley, K. J. (1986), "Likelihood Ratio Tests for a Change in the Multivariate Normal Mean," *J. American Statistical Association*, 81, 199–204.

[66] Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *Ann. Statist.*, 9, 1135–1151.

[67] Subbaiah, P., and Mudholkar, G. S. (1978), "A Comparison of Two Tests for the Significance of a Mean Vector," *J. American Statistical Association*, 73, 414–418.

[68] Sun, B., Zhou, S., and Shi. J. (2003), "An SPC Monitoring System for Cycle-Based Process Signals Using Wavelet Transform," Unpublished manuscript.

[69] Szu, H.H., Telfer, B.A., and Kadambe, S. (1992), "Neural Network Adaptive Wavelets for Signal Representation and Classification," Optical Engineering, 31, 1907–1916.

[70] Tang, D. I. (1994), "Uniformly More Powerful Tests in a One-Sided Multivariate Problem," *J. American Statistical Association*, 89, 1006–1011.

[71] Telfer, B.A., Szu, H.H., Dobeck, G.J., Garcia, J.P., Ko, H., Dubey, A., and Witherspoon, N. (1994), "Adaptive Wavelet Classification of Acoustic and Backscatter and Imagery," Optical Engineering, 33, 2192–2203.

[72] Tou, J., and Gonzalez, R. C. (1974), *Pattern Recognition Principles*, Addison-Wesley Publishing Company, Massachusetts.

[73] Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York: John Wiley & Sons, New York, NY.

[74] Vidakovic, B. (2000), "Unbalancing Data With Wavelet Transformations," *Technical Report*, Department of Statistics, Duke University, Durham, North Carolina.

[75] Vidakovic, B. (2001), "Wavelet-Based Functional Data Analysis: Theory, Applications and Ramifications," *Proceedings of PSFVIP-3*, March 18–20, Maui, Hawai, U.S.A..

[76] Wade, M. R. and Woodall, W. H. (1993), "A Review and Analysis of Cause-Selected Control Charts," *Journal of Quality Technology*, 25, 161–169.

[77] Walker, E., and Wright, S. P. (2002), "Comparing Curves Using Additive Models," *Journal of Quality Technology*, 34, 118–129.

[78] Wang, X. Z., Chen, B. H., Yang, S. H., and McGreavy, C. (1999), "Application of wavelets and Neural Networks to Diagnostic System Development, 2, An integrated Framework and its Application," *Computers and Chemical Engineering*, 23, 945–954.

[79] Wang, Y., and McDermott, M. P. (1998), "A Conditional Test for a Non-negative Mean Vector Based on a Hotelling's T^2-type Statistic," *J. Multivariate Analysis*, 65, 64–70.

[80] Weyrich, N. and Warhola, G. T. (1998), "Wavelet Shrinkage and Generalized Cross Validation for Image Denoising," *IEEE Transactions on Image Processing*, 7(1), 82-90.

[81] Woodall, W. H. (2000), "Controversies and Contradictions in Statistical Process Control," *Journal of Quality Technology*, 32, 341–349.

[82] Woodall, W. H., Spitzner, D. J., Montgomery, D. C., and Gupta, S. (2003), "Using Control Charts to Monitor Process and Product Profiles," submitted to *Journal of Quality Technology*.

[83] Zhou, W. (1998), "Structured Wavelet Antenna Signal Modeling and Random Scale Generalized Linear Model ", Ph.D thesis, Department of Statistics, North Carolina Sate University.

VITA

Myong-Kee Jeong was born November 6, 1968, in Korea. He received a B.S. in Industrial Engineering from HanYang University, Seoul, Korea, in 1991. He received the M.S. in Industrial Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Taejon, Korea, in 1993. From 1993 to 1999, he worked as a senior researcher at the Electronics and Telecommunications Research Institute (ETRI), Korea, where he developed an information system to provide design engineers a computerized reliability design and evaluation tool. At ETRI, he registered one patent: The Method of Estimating the Cut-off Connection Rate in the ATM Switching System Using Simulation (Patent Registration No. 1002373970000, Korea) in 2000. Since fall 1999, he has worked as a graduate research assistant while pursuing a doctoral degree in the School of Industrial and Systems Engineering at the Georgia Institute of Technology, Atlanta. His research interests include data mining in manufacturing and design, development of wavelet-based data reduction tools for decision-making with massive data, and process design, modeling and optimization in electronics and manufacturing processes. He taught a required course, Methods of Quality Improvement in the School of Industrial and Systems Engineering at Georgia Institute of Technology in summer and fall 2002. He was awarded a Freund International scholarship (2002-2004), which is a national competition award and was selected as the Face of Quality progress in December 2002. In 2001 and 2002, he received NAFSA Awards for Excellent International Students from the Association of International Educators. He was awarded the ETRI championship at a doubles match in fall 1997 and fall 1998. He has served since 2002 as the president of the Korean Graduate Student Bible Study at New Seoul Baptist Church